

R. W. HAMMING
Bell Telephone Laboratories

NUMERICAL METHODS FOR SCIENTISTS AND ENGINEERS

MC GRAW-HILL BOOK COMPANY, INC.
NEW YORK, SAN FRANCISCO, TORONTO, LONDON
1962

Физико-
Математическая
Библиотека
Инженера

Р. В. ХЕММИНГ

ЧИСЛЕННЫЕ МЕТОДЫ

для научных работников и инженеров

Перевод с английского
В. Л. АРЛАЗАРОВА, Г. С. РАЗИНОЙ
и А. В. УСКОВА

под редакцией
Р. С. ГУТЕРА

ИЗДАНИЕ ВТОРОЕ, ИСПРАВЛЕННОЕ

ИЗДАТЕЛЬСТВО «НАУКА»
ГЛАВНАЯ РЕДАКЦИЯ
ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ
МОСКВА 1972

Численные методы для научных работников и инженеров.
Хемминг Р. В.

Книга посвящена численным методам математического анализа, используемым на современных электронных вычислительных машинах. Она состоит из четырех частей.

Часть I, Дискретное исчисление конечных разностей (гл. 1-6), излагает основные понятия конечных разностей, суммирования конечных числовых рядов и конечных рядов Фурье.

Часть II, Приближение многочленами (гл. 7-20), содержит изложение классических численных методов интерполяции, численного интегрирования и численного решения дифференциальных уравнений, основанных на аппроксимации функции обычными алгебраическими многочленами. При этом рассматриваются приближения в смысле точного совпадения в узлах, в смысле наименьших квадратов и в смысле наименьшего отклонения по Чебышеву.

Часть III, Немногочленные приближения (гл. 21-27), посвящена аппроксимации функций с помощью экспоненциальных, а также с помощью рядов и интеграла Фурье.

Часть IV, Алгоритмы и эвристические методы (гл. 28-32), кроме некоторых известных алгоритмов для отыскания корней функции и для ряда задач линейной алгебры, рассматривает примеры моделирования, применения метода Монте-Карло и некоторые игровые задачи. Отдельная заключительная глава посвящена вопросам организации вычислительной работы.

Третья и четвертая части книги содержат ряд новых задач и методов. Изложение всех численных методов сопровождается разбором примеров из вычислительной практики автора.

Таблиц 32, рисунков 43, библиографических ссылок 44.

Второе издание печатается с матриц предыдущего, исправлены лишь замеченные опечатки.

Р. В. Хемминг

ЧИСЛЕННЫЕ МЕТОДЫ

(для научных работников и инженеров)

М., 1972 г., 400 стр. с илл.

Редактор Г. Я. Пирогова

Техн. редактор С. Я. Шкляр

Корректоры М. Ф. Алексеева, М. Л. Липелис

Печать с матриц. Подписано к печати 5/1 1972 г. Бумага 60×901/16. Физ. печ. л. 25.

Услови. печ. л. 25. Уч.-изд. л. 23,02. Тираж 27000 экз.

Цена книги 1 р. 83 к. Заказ № 18.

Издательство «Наука».

Главная редакция физико-математической литературы.

117071, Москва, В-71, Ленинский проспект, 15.

Ордена Трудового Красного Знамени Московская типография № 7 «Искра революции»

Главполиграфпрома Комитета по печати при Совете Министров СССР.

г. Москва, Трехпрудный пер., 9.

Предисловие редактора перевода	12
Из предисловия автора	14

ЧАСТЬ I

ДИСКРЕТНОЕ ИСЧИСЛЕНИЕ КОНЕЧНЫХ РАЗНОСТЕЙ

Глава 1. Исчисление разностей	17
§ 1.1. Введение и система обозначений	17
§ 1.2. Разностный оператор	19
§ 1.3. Повторные разности	21
§ 1.4. Таблицы разностей	23
§ 1.5. Факториалы	27
§ 1.6. Деление многочленов	29
§ 1.7. Числа Стирлинга первого рода	32
§ 1.8. Числа Стирлинга второго рода	34
§ 1.9. Пример	35
§ 1.10. Альтернативные замечания	36
§ 1.11. Общие замечания и справки	37
Глава 2. Погрешности округления	37
§ 2.1. Введение	37
§ 2.2. Область ответа	38
§ 2.3. Двойная точность	39
§ 2.4. Счет со значащими разрядами	39
§ 2.5. Статистический подход	40
§ 2.6. Случайное округление	41
§ 2.7. Переменная точность	41
§ 2.8. Оценка шума в таблице	41
§ 2.9. Теория «младшего значащего разряда»	47
§ 2.10. Теория «старшего значащего разряда»	49
§ 2.11. Анализ распространения ошибки при небольшом вычислении	52
§ 2.12. Общие замечания и библиография	53
Глава 3. Исчисление сумм	53
§ 3.1. Введение и система обозначений	53
§ 3.2. Формулы суммирования	56
§ 3.3. Суммирование по частям	58
§ 3.4. Общие замечания	59
Глава 4. Вычисление бесконечных рядов	59
§ 4.1. Введение	59
§ 4.2. Метод Куммера	61

§ 4.3. Некоторые специальные суммы	62
§ 4.4. Метод Эйлера	62
§ 4.5. Нелинейное преобразование	66
§ 4.6. Степенные ряды	67
§ 4.7. Разложение по специальным функциям	68
§ 4.8. Интегралы как приближения сумм	68
§ 4.9. Дигамма-функция	69
Глава 5. Уравнения в конечных разностях	71
§ 5.1. Система обозначений	71
§ 5.2. Пример разностного уравнения первого порядка	72
§ 5.3. Пример уравнения второго порядка	74
§ 5.4. Линейные разностные уравнения с постоянными коэффициентами	75
§ 5.5. Пример	76
Глава 6. Конечные ряды Фурье	78
§ 6.1. Введение	78
§ 6.2. Ортогональность на дискретном множестве точек	79
§ 6.3. Точность разложения	81
§ 6.4. Вычисление коэффициентов	83
§ 6.5. Метод двенадцати ординат	85
§ 6.6. Методы с минимумом умножений	87
§ 6.7. Разложение по косинусам	87
§ 6.8. Локальные ряды Фурье	88
 часть II	
ПРИБЛИЖЕНИЕ МНОГОЧЛЕНАМИ — КЛАССИЧЕСКИЙ ЧИСЛЕННЫЙ АНАЛИЗ	
Глава 7. Введение в многочленные приближения	90
§ 7.1. Ориентация	90
§ 7.2. Альтернативные формулировки	92
§ 7.3. Узловые точки, информация	95
§ 7.4. Класс функций	96
§ 7.5. Согласие	97
§ 7.6. Точность	98
Глава 8. Интерполяция многочленами. Данные с произвольными промежутками	99
§ 8.1. Философия	99
§ 8.2. Интерполяционные многочлены	99
§ 8.3. Метод интерполяции Лагранжа	103
§ 8.4. Интерполяционная формула Ньютона	106
§ 8.5. Другая форма для таблицы разделенных разностей	109
§ 8.6. Погрешность многочленной аппроксимации	110
§ 8.7. Трудности приближения многочленом	113
§ 8.8. О выборе узловых точек	116
Глава 9. Интерполяция многочленами. Равноотстоящие узлы	117
§ 9.1. Формула Ньютона для интерполирования	117
§ 9.2. Интерполирование в таблицах	118
§ 9.3. Ромбовидная диаграмма	119
§ 9.4. Замечания к выведенным формулам	123
§ 9.5. Смешанные интерполяционные формулы	124

Глава 10. Единый метод нахождения интерполяционных формул	125
§ 10.1. Введение	125
§ 10.2. Несколько типичных формул интегрирования	127
§ 10.3. Фиксированные узлы	132
§ 10.4. Некоторые примеры формул	135
§ 10.5. Значения функции и производной в фиксированных точках	137
§ 10.6. Свободные узлы; квадратура Гаусса	139
§ 10.7. Смешанный случай	141
§ 10.8. Замечания	142
§ 10.9. Линейные ограничения на веса	144
§ 10.10. Формула Грегори	147
§ 10.11. Выводы	150
Глава 11. О нахождении остаточного члена формулы	152
§ 11.1. Потребность в остаточном члене	152
§ 11.2. Порядок остаточного члена	152
§ 11.3. Функция влияния	153
§ 11.4. Случай, когда $G(s)$ имеет постоянный знак	156
§ 11.5. Случай, когда функция влияния меняет знак	158
§ 11.6. Слабое место в методе рядов Тейлора	160
Глава 12. Формулы для определенных интегралов	161
§ 12.1. Введение	161
§ 12.2. Формулы Ньютона—Котеса	164
§ 12.3. Использование формулы Грегори	166
§ 12.4. Открытые формулы	168
§ 12.5. Квадратура Гаусса	169
§ 12.6. Формулы интегрирования смешанного гауссового типа	170
§ 12.7. Суммирование рядов	171
§ 12.8. Эффекты замены переменной	172
§ 12.9. Интегралы с параметром	173
Глава 13. Неопределенные интегралы	173
§ 13.1. Описание содержания главы и система обозначений	173
§ 13.2. Несколько простых формул для неопределенных интегралов	175
§ 13.3. Общий метод	177
§ 13.4. Ошибка вследствие отбрасывания членов	178
§ 13.5. Устойчивость	181
§ 13.6. Шум округления	184
§ 13.7. Итоги	186
§ 13.8. Некоторые общие замечания	187
§ 13.9. Экспериментальная проверка устойчивости	189
§ 13.10. Пример интеграла свертки, иллюстрирующий идею устойчивости	189
Глава 14. Введение в дифференциальные уравнения	191
§ 14.1. Природа и смысл дифференциальных уравнений	191
§ 14.2. Поле направлений	192
§ 14.3. Численное решение	193
§ 14.4. Пример	195
§ 14.5. Устойчивость метода простого прогноза	197
§ 14.6. Устойчивость коррекции	198
§ 14.7. Несколько общих замечаний	200
§ 14.8. Системы уравнений	201

Глава 15. Общая теория методов прогноза и коррекции	202
§ 15.1. Введение	202
§ 15.2. Ошибка от отбрасывания членов	204
§ 15.3. Устойчивость	205
§ 15.4. Помехи округления	209
§ 15.5. Прогноз по трем точкам	209
§ 15.6. Прогнозы типа Милна	210
§ 15.7. Прогнозы типа Адамса—Башфорта	212
§ 15.8. Общие замечания о выборе метода	213
§ 15.9. Выбор прогноза	214
§ 15.10. Некоторые формулы	215
§ 15.11. Выбор шага и оценка точности	216
§ 15.12. Экспериментальная проверка	219
Глава 16. Специальные методы интегрирования обыкновенных дифференциальных уравнений	220
§ 16.1. Введение и общее описание	220
§ 16.2. Методы Рунге—Кутты	221
§ 16.3. Методы для уравнения второго порядка, когда отсутствует y'	222
§ 16.4. Линейные уравнения	224
§ 16.5. Метод, который использует значения y , y' и y''	225
§ 16.6. Случай, когда решение трудно аппроксимировать многочленом	226
§ 16.7. Краевые задачи	229
Глава 17. Метод наименьших квадратов. Теория	232
§ 17.1. Введение	232
§ 17.2. Метод наименьших квадратов	232
§ 17.3. Другие критерии	234
§ 17.4. Ошибки с нормальным распределением	234
§ 17.5. Проведение подходящего многочлена	237
§ 17.6. Ортогональные функции	240
§ 17.7. Общие свойства ортогональных функций	242
§ 17.8. Неравенство Бесселя и полнота	244
§ 17.9. Метод наименьших квадратов и коэффициенты Фурье	245
§ 17.10. Ортогональные многочлены	247
§ 17.11. Классические ортогональные многочлены	249
§ 17.12. Сравнение метода наименьших квадратов и разложения в степенные ряды	250
§ 17.13. Метод наименьших квадратов с ограничениями; продолжение примера из § 1.9	251
§ 17.14. Последние замечания о методе наименьших квадратов	252
Глава 18. Метод наименьших квадратов. Практика	252
§ 18.1. Общие замечания о многочленном случае	252
§ 18.2. Трехчленное рекуррентное соотношение	253
§ 18.3. Построение квазиортогональных многочленов	255
§ 18.4. Немногочленный случай	255
§ 18.5. Нелинейные параметры	256
Глава 19. Многочлены Чебышева	257
§ 19.1. Введение	257
§ 19.2. Некоторые тождества	259
§ 19.3. Критерий Чебышева	260
§ 19.4. Экономизация	262

§ 19.5. Механизация процесса экономизации	263
§ 19.6. Смещенные многочлены Чебышева	265
§ 19.7. τ -процесс Ланцоша	266
§ 19.8. Видоизменение τ -метода	268
§ 19.9. Несколько замечаний о чебышевском приближении	270
§ 19.10. Критерий совпадения моментов	270
Глава 20. Рациональные функции	272
§ 20.1. Введение	272
§ 20.2. Непосредственный подход	273
§ 20.3. Чебышевское приближение рациональными функциями	274
§ 20.4. Обратные разности (симметричные)	275
§ 20.5. Пример	278
часть III	
НЕМНОГОЧЛЕННЫЕ ПРИБЛИЖЕНИЯ	
Глава 21. Периодические функции. Аппроксимация Фурье	280
§ 21.1. Цель этой теории	280
§ 21.2. Замена переменных и выбор узлов	281
§ 21.3. Ряды Фурье; периодические явления	282
§ 21.4. Интерполяция периодических функций	285
§ 21.5. Интегрирование	288
§ 21.6. Метод общего оператора	290
§ 21.7. Несколько замечаний относительно общего метода	293
Глава 22. Сходимость рядов Фурье	294
§ 22.1. Сходимость степенных рядов и рядов Фурье	294
§ 22.2. Функции с простым разрывом	295
§ 22.3. Функция, имеющая непрерывные производные более высокого порядка	297
§ 22.4. Улучшение сходимости ряда Фурье	298
§ 22.5. Спектр мощности	299
§ 22.6. Явление Гиббса	300
§ 22.7. Сигма-множители Ланцоша	301
§ 22.8. Сравнение методов сходимости	303
§ 22.9. Техника дифференцирования по Ланцошу	304
Глава 23. Непериодические функции. Интеграл Фурье	305
§ 23.1. Цель главы	305
§ 23.2. Обозначения и краткое изложение результатов	306
§ 23.3. Интеграл Фурье	310
§ 23.4. Преобразование Фурье некоторых функций	311
§ 23.5. Функции с ограниченным спектром и теорема выборки	313
§ 23.6. Теорема свертки	315
§ 23.7. Эффект конечного суммирования	316
Глава 24. Линейные фильтры. Сглаживание и дифференцирование	317
§ 24.1. Введение	317
§ 24.2. Пример простого сглаживающего фильтра	318
§ 24.3. Пример построения фильтра	319
§ 24.4. Фильтры вообще	320
§ 24.5. Анализ простых формул для дифференцирования	321
§ 24.6. Как избежать вычисления производных?	322

§ 24.7. Метод Филона	323
§ 24.8. Заключительные замечания	325
Глава 25. Интегралы и дифференциальные уравнения	326
§ 25.1. Содержание главы	326
§ 25.2. Метод передаточной функции для интегрирования	327
§ 25.3. Общие формулы интегрирования	331
§ 25.4. Дифференциальные уравнения	332
§ 25.5. Построение фильтров по методу Чебышева	334
§ 25.6. Некоторые детали метода Чебышева	336
Глава 26. Экспоненциальная аппроксимация	340
§ 26.1. Введение	340
§ 26.2. О нахождении формул, использующих экспоненты, когда показатели экспонент известны	340
§ 26.3. Неизвестные показатели	342
§ 26.4. Предупреждения	343
§ 26.5. Экспоненты и многочлены	344
§ 26.6. Остаточные члены	344
Глава 27. Особенности	344
§ 27.1. Введение	344
§ 27.2. Пример интеграла с особенностью в бесконечности	345
§ 27.3. Особенность в линейном дифференциальном уравнении	346
§ 27.4. Общие замечания	349
часть IV	
АЛГОРИТМЫ И ЭВРИСТИЧЕСКИЕ МЕТОДЫ	
Глава 28. Нахождение нулей	350
§ 28.1. Алгоритмы и эвристические методы	350
§ 28.2. Метод деления пополам для нахождения корня функции	351
§ 28.3. Линейная интерполяция	352
§ 28.4. Параболическая интерполяция	352
§ 28.5. Некоторые общие замечания	353
§ 28.6. Метод Берстоу для нахождения комплексных корней мно- гочлена	355
Глава 29. Системы линейных алгебраических уравнений	359
§ 29.1. Введение	359
§ 29.2. Метод исключения Гаусса	360
§ 29.3. Варианты метода Гаусса	362
§ 29.4. Метод Гаусса—Зайделя	363
§ 29.5. Повышенная точность	364
§ 29.6. Общие замечания	364
Глава 30. Обращение матриц и собственные значения	365
§ 30.1. Введение	365
§ 30.2. Обращение матрицы методом исключения по Гауссу	365
§ 30.3. Задача нахождения собственных значений	366
§ 30.4. Наименьшие собственные значения	368
§ 30.5. Несколько замечаний	368
Глава 31. Некоторые примеры моделирования	369
§ 31.1. Введение	369
§ 31.2. Простой пример дискретного моделирования	370

§ 31.3. Пример моделирования складских операций	374
§ 31.4. Трехмерные крестики — нолики	375
§ 31.5. Общие замечания о дискретном моделировании	379
§ 31.6. Непрерывное моделирование	380
Глава 32. Случайные числа и методы Монте-Карло	381
§ 32.1. Понятие случайного числа	381
§ 32.2. Генерирование случайных чисел в машине, работающей в двоичной системе	382
§ 32.3. Генерирование случайных чисел на десятичной машине	386
§ 32.4. Другие распределения	386
§ 32.5. Метод Монте-Карло	388
§ 32.6. Еще одна иллюстрация метода Монте-Карло	389
§ 32.7. Метод жулика	390
Глава N+1. Искусство вычислять для инженеров и ученых	391
§ N+1.1. Важность вопроса	391
§ N+1.2. Что мы собираемся делать с ответом?	392
§ N+1.3. Что мы знаем?	393
§ N+1.4. Обдумывание вычислений	394
§ N+1.5. Повторение предыдущих шагов	395
§ N+1.6. Оценка усилий, необходимых для решения задачи	395
§ N+1.7. Изменения первоначального плана	396
§ N+1.8. Философия	397
§ N+1.9. Заключительные замечания	398
Литература	399

ПРЕДИСЛОВИЕ РЕДАКТОРА ПЕРЕВОДА

Имя Р. В. Хемминга — известного американского ученого, бывшего президента ассоциации по вычислительным машинам, руководителя математической службы «Bell Telephone Laboratories» — и его работы в области вычислительной математики и теории информации достаточно хорошо известны и не нуждаются в особых рекомендациях. Трудно, однако, удержаться от использования предоставившейся возможности рекомендовать читателю замечательную книгу.

Книга «Численные методы для научных работников и инженеров» бесспорно является выдающимся явлением в математической литературе. Она удивительным образом сочетает широту охватываемого материала, глубину подхода к нему и практичность в лучшем смысле этого слова, нигде не переходящую в узкий практицизм.

Среди уже довольно многочисленных книг по вычислительной математике книга Р. В. Хемминга выделяется и по содержанию и по форме.

Прежде всего, в ней нашли широкое и полное отражение идеи П. Л. Чебышева. Не только в зарубежной, но и в нашей русской литературе многочисленные аспекты чебышевских идей и методов численного анализа не получали еще столь полного и широкого освещения. Другой особенностью книги, относящейся к ее содержанию, является большое внимание, уделяемое различного рода немногочленным приближениям. В книге достаточно подробно рассмотрена аппроксимация функции рациональными и экспоненциальными, а также функциями с ограниченным спектром. Последнее особенно интересно и имеет большое значение для практики применения численных методов.

Указанные особенности содержания легко объясняются заметным впечатком, наложенным на него личными научными интересами автора, и являются следствием единого и нового подхода к вычислительной математике — с точки зрения теории информации. Эта точка зрения проводится систематически, и ее преимущества будут легко замечены читателями.

Еще больший отпечаток наложили личные научные интересы и вкусы автора на форму изложения, стиль и манеру письма. Книга написана весьма субъективно, и от этого интерес к ней особенно возрастает.

Если верить бытующей «классификации», согласно которой работающие в области вычислительной математики делятся на тех, кто доказывает сходимость вычислительных процессов и существование решений, и тех, кто применяет вычислительные процессы и получает решения, то Р. В. Хемминг является видным представителем вычислителей второго из этих типов. Огромный личный опыт вычислителя не позволяет ему ограничиваться беспристрастным изложением того или иного вычислительного метода, не освещая своего отношения к нему. Многие методы иллюстрируются автором примерами из его собственной вычислительной практики. Эти качества особенно важны для книги по вычислительной математике, где «искусство вычислителя» и «маленькие хитрости» лишь в редких случаях не позволяют уменьшить вычислительную работу в сотню-другую раз или увеличить точность во столько же.

При изложении вычислительных методов автор уделяет большое внимание физической сущности рассматриваемых математических задач. В основу всей книги положены два тезиса, неоднократно повторяемых. Это

«цель расчетов — понимание, а не числа»

и

«прежде чем решать задачу, подумай, что делать с ее решением».

Большой интерес представляет ($N+1$)-я глава книги, посвященная вопросам организации вычислительной работы и взаимоотношениям вычислителей с заказчиками. Как справедливо отмечает автор, на эти темы писать не принято, несмотря на всю их практическую важность. Эта глава, разумеется, наиболее субъективна, и легко представить читателей, не разделяющих высказываемых автором воззрений. Для них хорошо процитировать лишь заключительный абзац, завершающий книгу. Впрочем, еще лучше, если читатель прочтет его в тексте.

При работе над переводом мы полностью сохранили структуру книги и стремились к тому, чтобы как можно точнее передать на русском языке текст и дух подлинника. Сохранена без изменения и библиография в конце книги; мы ограничились лишь указаниями на то, какие из цитируемых автором книг имеются в русском переводе. Несколько замечаний, которые считал возможным сделать редактор, относятся главным образом к согласованию терминологии или дополнительным литературным ссылкам. Все они вынесены в подстрочные примечания и их принадлежность редактору всюду оговорена.

Мы надеемся, что книга Р. В. Хемминга будет по достоинству оценена читателями. У нее есть все основания стать настольной книгой для всех, кто занимается вычислительной работой или связан с нею — от руководителей институтов и отделов до квалифицированных лаборантов.

Эта книга написана для научных работников и инженеров, которые собираются использовать современные цифровые вычислительные машины как средство для своих исследований. Она может также служить первоначальным учебником численного анализа. Автор убежден, что таким читателям, для того чтобы они могли понять, какое отношение имеют полученные на машине результаты к их проблемам, нужен не справочник и не сводка отдельных результатов, а скорее связное изложение основных идей вычислительной математики. Как утверждает девиз этой книги, мы ищем смысл, а не числа.

Книга отличается от имеющихся по следующим пунктам:

1. Есть много прекрасных книг, написанных с точки зрения людей, пользующихся арифмометрами; в этой книге предполагается, что будет использована большая цифровая вычислительная машина. Различие здесь не в том, что можно работать с большими задачами, а в том, что появляется совершенно другой подход к ним.

2. Имеется ряд очень хороших книг, которые являются собранием несвязанных глав (часто написанных разными авторами) и которые не в состоянии дать единое представление об области в целом. Одна из главных целей этой книги — показать, как можно объединить разные частные результаты в рамках общих идей и методов. Таким образом, читатель может надеяться понять отношение между некоторыми из многих различных формул для одной и той же цели. Он сможет также выводить много новых формул для удовлетворения своих требований.

3. Существует ряд хороших книг, написанных математиками для математиков. В этой книге мы старались подать материал в форме, удобной для тех, кто больше заинтересован в использовании новых мощных вычислительных средств, чем в красоте вывода формул или в дальнейших исследованиях.

4. В большинстве книг преобладает использование для численных методов полиномиальных приближений. При таком подходе остаточный член обычно выражается через производную высокого порядка, которую редко удается оценить даже и грубо.

В нашей книге используется метод функций с ограниченным спектром, который хорошо известен электротехникам, но мало исполь-

зуется в вычислительной математике. В этом случае ошибка выражается через саму функцию. Модель с ограниченным спектром легче применять к физическим задачам, чем полиномиальную модель, и поэтому первая значительно облегчает планирование и интерпретацию вычислений. Метод функций с ограниченным спектром связывает также вычисления с теоремой выборки теории информации и с рядом других теорий.

5. В литературе по численному анализу существует тенденция уделять основное внимание решению систем линейных алгебраических уравнений, обращению матриц, нахождению собственных значений матриц и корней многочленов. На практике потребитель должен получить от вычислительного центра не только подпрограммы вычисления элементарных функций, но также и подпрограммы указанных выше вычислительных процессов. Поэтому, если потребитель может при обращении к такой подпрограмме потребовать выполнения соответствующего его задаче критерия и понять полученные результаты (а также при условии, что подпрограммы не съедают слишком много машинного времени), то ему не так уж важно, как эти подпрограммы составлены.

В самом деле, сомнительно, чтобы потребитель захотел решать вырожденную систему линейных уравнений. Скорее он захочет узнать, отчего она оказалась вырожденной, чтобы понять, что неверно в постановке задачи. Поэтому здесь эти вычислительные методы обсуждаются ровно настолько, сколько требуется, чтобы читатель понял, на что нужно обратить внимание и чего можно ожидать. Предполагается, что если встретятся затруднения, то он обратится к соответствующему специалисту.

6. Книга не содержит числовых примеров с многими десятичными знаками. Автор намеревался привести много табулограмм, но постепенно понял две вещи.

Во-первых, простые примеры, просчитанные с небольшой точностью, которые читатель может проследить глазами или, в крайнем случае, с логарифмической линейкой, много более поучительны, чем десятизначные вычисления, на которые читатель лишь взглянет и сразу перевернет страницу.

Во-вторых, в наши дни большая часть планирования и программирования задач делается в действительности при полном отсутствии рабочих примеров. Иногда возражают, что тем больше имеется причин приводить в книге много тщательно разобранных примеров.

Автор не согласен с этим возражением. Если придется разрабатывать план и писать программу без каких-либо числовых примеров, то чем скорее мы усвоим это и научимся обходиться без примеров, тем лучше. Конечно, очень полезно иметь за плечами год или два опыта решения задач на арифмометре, но это требует больших потерь времени.

В любой книге, задуманной как учебник годовичного курса, приходится опускать много полезного материала. В результате в этой книге отсутствует ряд вещей:

1. В книге ничего не говорится о программировании. Предполагается, что читатель знаком с каким-либо языком программирования, например ФОРТРАН или АЛГОЛ.

2. Ради сохранения разумного объема книги пришлось исключить функции нескольких переменных. Поэтому в книгу не включена такая важная тема, как дифференциальные уравнения в частных производных. Это серьезное упущение, но мы можем отослать читателя к другим книгам, например Вазова и Форсайта [9].

3. Как уже указывалось, решение систем линейных алгебраических уравнений, обращение матриц, отыскание собственных значений матриц и корней многочленов рассмотрены лишь настолько, чтобы дать читателю представление, что можно сделать, чего можно ожидать и что следует искать.

4. Рассмотрение алгоритмов вообще явно недостаточно; но разумная теория только начинает появляться в литературе. В настоящем состоянии хаоса подробное изложение темы не кажется подходящим для первоначального курса.

5. Эта книга написана за один год, когда автор читал такой курс в Стенфордском университете. Большая часть материала приводится не по источникам, а по беглым записям и на основании пятнадцатилетней вычислительной практики, когда нас интересовало главным образом получение результатов. Поэтому новые и неизвестные факты перемешаны в книге со старым и хорошо известным.

Материал книги распределен следующим образом. Первая часть охватывает дискретное исчисление конечных разностей, являющееся основой большинства численных методов. В ней не возникает вопрос о пределах ошибок аппроксимации. Во второй части предполагается, что по узловым точкам проведен интерполяционный многочлен. В ней содержится то, что можно назвать классической частью численного анализа. Материал третьей части основан на предположении, что между узловыми точками функция аппроксимируется функцией с ограниченным спектром. Эта часть связывает многие разделы вычислительной математики с другими областями, как теорема выборки теории информации, проектирование фильтров и передаточная функция в электротехнике и т. п. В третьей части также кратко изложены вопросы приближения экспонентами и работа с особенностями. Четвертая часть начинается с алгоритмов, рассматривает эвристические методы и случайные процессы и кончается главой об искусстве вычислений.

Не следует считать, что автору принадлежит все то новое, что может встретиться читателю в этой книге. Тем не менее автор полностью берет на себя ответственность за изложение и возможные ошибки.

ДИСКРЕТНОЕ ИСЧИСЛЕНИЕ КОНЕЧНЫХ РАЗНОСТЕЙ

ГЛАВА I

ИСЧИСЛЕНИЕ РАЗНОСТЕЙ

§ 1.1. Введение и система обозначений

Первая часть книги посвящается изучению функций одного переменного, определенных на дискретном множестве равноотстоящих точек. Например,

$$f(a), \quad f(a+h), \quad f(a+2h), \quad \dots, \quad f[a+(n-1)h]$$

есть множество значений функции $f(x)$ в n равноотстоящих точках $x=a, a+h, \dots, a+(n-1)h$.

В качестве другого примера такой функции можно указать последовательность частичных сумм ряда

$$S(r) = \sum_{k=1}^r a_k \quad (r=1, 2, \dots, n).$$

В первой части всюду, кроме гл. 4, в которой идет речь о бесконечных рядах, число значений функции конечно. Поэтому вопросы существования и ограниченности здесь обычно не представляют трудностей; легко видеть, выполнены или не выполнены те или иные условия. Ограничившись такими простыми случаями, можно рассмотреть большую часть методов численного анализа только в конечноразностных соотношениях, не касаясь более сложных вопросов.

Поскольку вычислительная машина может выполнить лишь конечное число операций, ясно, что предельный переход непосредственно машиной выполнен быть не может. В связи с этим в численном анализе приходится заниматься вопросами получения оценок результатов предельных переходов при помощи конечного числа действий.

Стоит вспомнить, что дифференциальные уравнения, встречающиеся на практике, обычно возникают из физических задач через посредство конечноразностных схем. Если обратный переход от непрерывного случая к дискретному невозможен, то имеются основания сомневаться в физическом смысле соответствующих уравнений.

Во второй части предполагается, что в промежутке между заданными точками (узлами) наша функция представляется или аппроксимируется многочленом или рациональной функцией; в третьей части рассмотрено представление функциями с ограниченным спектром*), суммой показательных функций или функциями какого-либо другого вида, определяемого характером особенностей. Во всех случаях, однако, мы будем оперировать со значениями функций на данном дискретном множестве точек.

Иногда бывает выгодно использовать множество неравноотстоящих узлов. Но даже и в этом случае идеи и формальные преобразования часто связаны с идеями и преобразованиями для равноотстоящих узлов. Таким образом, методы, используемые в конечноразностных схемах с равноотстоящими точками являются основой большей части численного анализа.

Первая часть книги существенно использует параллели между исчислением разностей и дифференциальным исчислением. Предполагается, что читатель знаком с дифференциальным исчислением. Все трудности, которые могут возникнуть у читателя в главах 1, 3 и 5, будут, вероятно, объясняться недостаточным пониманием соответствующего материала в дифференциальном исчислении.

Большинство современных вычислительных машин работает в режиме «плавающей запятой». Для десятичной системы счисления это означает, что числа пишутся в виде

$$\pi = 0,31415927 \cdot 10^1, \quad 10\pi = 0,31415927 \cdot 10^2, \quad \pi/_{10} = 0,31415927 \cdot 10^0,$$

и т. д. Это — общепринятое обозначение**), и мы всегда будем предполагать, что вычисления ведутся с плавающей запятой, если только не оговаривается противное.

В этой книге предполагается, что все вычисления производятся на универсальной вычислительной машине и что вычислительным центром обеспечены: удолетворительная система программирования, удовлетворительная система контроля и разумная библиотека подпрограмм. Единственным ручным вычислением предполагается расчет, необходимый, чтобы убедиться, что данная программа работает надлежащим образом, а также счет, требуемый для оценки времени решения на вычислительной машине.

*) Термин «функция с ограниченным спектром» есть техническое выражение, приблизительно означающее, что у функции существуют частоты только в данной полосе. Так, функция, имеющая частоты в полосе от 1 до 10 периодов в секунду, описывает определенный класс функций с ограниченным спектром. Точное определение будет дано в соответствующем месте книги (см. гл. 23).

**) Во многих машинах показатель степени (порядок) записывается различными условными способами; в частности, например, к показателю степени прибавляют те или иные константы. Для наших целей это не имеет никакого значения.

Замена переменного

$$x_k = a + kh$$

в первом примере этого параграфа новой переменной

$$t_k = \frac{x_k - a}{h} = k \quad (k = 0, 1, \dots, n-1) \quad (1.1-1)$$

приводит рассмотренный случай к удобной стандартной форме. Благодаря этому в первой части, если не сделано никаких оговорок, выражение $f(x)$ будет рассматриваться для значений $x = 0, 1, \dots, n-1$. Иногда удобно пользоваться обозначениями

$$f(0) = f_0,$$

$$f(1) = f_1,$$

$$\dots$$

С точки зрения результата вычислений обычно бывает безразлично, изменим ли мы данную задачу в соответствии с единицей измерения, или же преобразуем разностное выражение в соответствии с шагом в задаче. Это аналогично преобразованию окружности

$$(x-a)^2 + (y-b)^2 = r^2$$

в аналитической геометрии заменой переменных

$$x-a = x', \quad y-b = y'$$

в окружность

$$x'^2 + y'^2 = r^2.$$

Это преобразование можно рассматривать с двух точек зрения. Можно считать, что изменяется координатная система, что приведет к новым координатам для тех же самых точек окружности. В другом случае можно принять, что окружность переходит в новое положение, тогда как координатная система остается неподвижной.

Упражнения

1.1-1. Пусть x принимает значения 11, 9, 7, ..., -11. Найти преобразование, которое приводит его к стандартной форме 0, 1, ..., 11.

О т в е т: $t = (-x + 11):2$.

1.1-2. Привести к стандартной форме $x = 3; 3,5; 4,0; 4,5; \dots; 10$.

О т в е т: $t = 2(x - 3)$.

§ 1.2. Разностный оператор

Основным оператором в исчислении конечных разностей является разностный оператор Δ , определенный равенством

$$\Delta f(x) = f(x+h) - f(x). \quad (1.2-1)$$

Этот оператор знаком читателю из анализа, где он используется при определении производной. Тесная связь между производной и

разностью является основой большинства применений конечноразностных выражений для приближения выражений из дифференциального исчисления.

Оператор Δ можно представлять как существующий отдельно и действующий на функцию $f(x)$ точно так же, как мы часто рассматриваем производную $\frac{df(x)}{dx}$ как оператор $\frac{d}{dx}$, действующий на $f(x)$, и интеграл $\int f(x) dx$ — как интегральный оператор $\int \dots dx$, действующий на функцию $f(x)$.

Разностный оператор линеен, как дифференциальный и интегральный, т. е. если a и b постоянны, то

$$\Delta [af(x) + bg(x)] = a \Delta f(x) + b \Delta g(x). \quad (1.2-2)$$

Линейное свойство оператора делает оператор Δ особенно простым для применения во многих случаях.

В качестве примера рассмотрим функцию

$$y = ax^2 + bx + c.$$

Используя (1.2-2), получим

$$\Delta y = a \Delta x^2 + b \Delta x + c \Delta 1 = a [(x+h)^2 - x^2] + b [(x+h) - x] + c [1 - 1] = 2ahx + ah^2 + bh. \quad (1.2-3)$$

Разностный оператор, действуя на произведение, дает

$$\begin{aligned} \Delta [f(x)g(x)] &= f(x+h)g(x+h) - f(x)g(x) = \\ &= f(x+h)g(x+h) - f(x+h)g(x) + f(x+h)g(x) - f(x)g(x) = \\ &= f(x+h)\Delta g(x) + g(x)\Delta f(x) \end{aligned}$$

или

$$\Delta [f(x)g(x)] = f(x)\Delta g(x) + g(x+h)\Delta f(x). \quad (1.2-4)$$

Если исключить то, что аргумент одной из функций есть $x+h$, формула (1.2-4) соответствует формуле дифференциального исчисления

$$\frac{d}{dx} [f(x)g(x)] = f(x)g'(x) + g(x)f'(x).$$

Аналогично для отношения функций имеем

$$\begin{aligned} \Delta \frac{f(x)}{g(x)} &= \frac{f(x+h)}{g(x+h)} - \frac{f(x)}{g(x)} = \\ &= \frac{f(x+h)g(x) - g(x+h)f(x) + [f(x)g(x) - f(x)g(x)]}{g(x)g(x+h)} = \\ &= \frac{g(x)\Delta f(x) - f(x)\Delta g(x)}{g(x)g(x+h)}, \quad (1.2-5) \end{aligned}$$

что опять напоминает соответствующую формулу для дифференциального исчисления, а именно:

$$\frac{d}{dx} \frac{f(x)}{g(x)} = \frac{g(x)f'(x) - f(x)g'(x)}{g^2(x)}.$$

В книгах по анализу обычно дается длинная таблица формул для производных. Аналогичная таблица используется и в исчислении разностей. Приведем следующую короткую таблицу:

$$\Delta \sin(ax + b) = 2 \sin \frac{ah}{2} \cos \left[a \left(x + \frac{h}{2} \right) + b \right], \quad (1.2-6)$$

$$\Delta \cos(ax + b) = -2 \sin \frac{ah}{2} \sin \left[a \left(x + \frac{h}{2} \right) + b \right], \quad (1.2-7)$$

$$\Delta \operatorname{tg}(ax + b) = \sin ah \sec(ax + b) \sec[a(x + h) + b], \quad (1.2-8)$$

$$\Delta a^x = a^x (a^h - 1), \quad (1.2-9)$$

$$\Delta 2^x = 2^x (2^h - 1), \quad (1.2-10)$$

$$\Delta \ln x = \ln \left(1 + \frac{h}{x} \right). \quad (1.2-11)$$

Роль, которую играет число e в дифференциальном исчислении, в исчислении конечных разностей в некотором отношении играет число 2. Именно, если $a^h = 2$, то $\Delta a^x = a^x$. В частности, если $a = 2$ и $h = 1$, то $\Delta 2^x = 2^x$.

Упражнение 1.2-1. Проверить равенства с (1.2-6) по (1.2-11).

§ 1.3. Повторные разности

Так как $\Delta f(x)$ есть функция от x , то, применив к ней снова оператор Δ , можно получить

$$\Delta [\Delta f(x)] = \Delta^2 f(x). \quad (1.3-1)$$

Это обозначение соответствует обозначению для второй производной в дифференциальном исчислении

$$\frac{d}{dx} \frac{df(x)}{dx} = \frac{d^2 f(x)}{(dx)^2} = \frac{d^2 f(x)}{dx^2}.$$

Вообще

$$\Delta^r f(x) = \Delta [\Delta^{r-1} f(x)]. \quad (1.3-2)$$

В примере

$$y(x) = ax^3 + bx + c$$

мы имели (уравнение (1.2-3))

$$\Delta y(x) = 2ahx + ah^2 + bh;$$

тогда

$$\Delta^2 y(x) = 2ah^2.$$

Продолжая тем же способом, получаем

$$\Delta^3 y(x) = 0.$$

Тот факт, что $\Delta^3 y(x) = 0$, — не случайность, а следствие важной теоремы:

Основная теорема исчисления разностей. Для многочлена степени n

$$y(x) = a_0 + a_1 x + a_2 x^2 + \dots + a_n x^n \quad (a_n \neq 0)$$

n -я разность постоянна и равна $a_n n! h^n$, $(n+1)$ -я разность равна нулю. Доказательство легче всего понять, рассмотрев предварительно такую лемму:

Лемма. Если $y(x)$ — многочлен степени n , то $\Delta y(x)$ есть многочлен степени $n-1$ (т. е. степень $\Delta y(x)$ ровно на единицу меньше, чем степень $y(x)$).

Доказательство леммы. Для функции $y(x) = x^n$, используя разложение по биному ^{*}, находим

$$\begin{aligned} \Delta y(x) &= (x+h)^n - x^n = \sum_{k=0}^n C(n, k) h^{n-k} x^k - x^n = \\ &= nhx^{n-1} + \frac{n(n-1)h^2}{2} x^{n-2} + \dots + h^n. \end{aligned}$$

Таким образом, при применении оператора Δ член x^n становится многочленом степени $n-1$ со старшим коэффициентом равным nh . Используя свойство линейности (уравнение (1.2-2)), находим, что оператор Δ уменьшает степень каждого члена в многочлене на единицу. Член $nx^{n-1}ha_n$ не может уничтожиться, следовательно, лемма доказана.

Доказательство теоремы. Применив к многочлену n -й степени доказанную лемму n раз, убедимся, что его n -я разность постоянна и коэффициент при a_n есть $n!h^n$. Следующие разности обращаются в нуль. Этим теорема доказана.

Приведенная теорема имеет большое значение в классической части численного анализа.

Упражнение 1.3-1. Полагая $h=1$, вычислить вторую и четвертую разности многочлена $y = x^4 - 4x^3 + 6x^2 - 4x + 1$.

О т в е т: $\Delta^2 y = 12x^2 + 2$; $\Delta^4 y = 24$.

^{*} Мы используем для биномиальных коэффициентов более старое обозначение $C(n, k)$ вместо популярного в настоящее время $\binom{n}{k}$ (или любого другого с верхним или нижним, или тем и другим вместе индексами), так как прежние обозначение легче набрать, оно может быть напечатано вычислительной машиной и лучше выглядит в середине предложения и формулы.

§ 1.4. Таблицы разностей

Если приходится использовать высокие разности, полезно представлять их расположенными в виде таблицы (см. таблицу 1.4-1, где принимается $h=1$), хотя, возможно, разности не расположены в таком порядке в машине.

Таблица 1.4-1

Таблица разностей

x	$y(x)$	$\Delta y(x)$	$\Delta^2 y(x)$	$\Delta^3 y(x)$
0	$y(0)$	$\Delta y(0)$		
1	$y(1)$	$\Delta y(1)$	$\Delta^2 y(0)$	$\Delta^3 y(0)$
2	$y(2)$	$\Delta y(2)$	$\Delta^2 y(1)$	$\Delta^3 y(1)$
3	$y(3)$	$\Delta y(3)$	$\Delta^2 y(2)$	$\Delta^3 y(2)$
4	$y(4)$	$\Delta y(4)$	$\Delta^2 y(3)$	\vdots
5	$y(5)$	\vdots	\vdots	\vdots
\vdots	\vdots	\vdots	\vdots	\vdots

Таблица 1.4-2

Таблица разностей для

$$\text{Si}(x) = \int_0^x \frac{\sin t}{t} dt$$

x	$\text{Si}(x)$	Δ	Δ^2	Δ^3
0,0	0,0000	999		
0,1	0,0999	997	-2	
0,2	0,1996	989	-8	-6
0,3	0,2985	980	-9	-1
0,4	0,3965	966	-14	-5
0,5	0,4931	950	-16	-2
0,6	0,5881	931	-19	-3
0,7	0,6812	909	-22	-3
0,8	0,7721	884	-25	-3
0,9	0,8605	856	-28	
1,0	0,9461			

В качестве примера таблицы разностей рассмотрим значения интегрального синуса (таблица 1.4-2). По общепринятым обозначениям разности пишутся так, как если бы запятая в десятичном числе следовала за последними разрядами, отведенными для значения функции. Чтобы проверить арифметику, заметим, что в таблице 1.4-2 сумма чисел в любом столбце разностей, прибавленная к верхнему числу соседнего столбца слева, равна нижнему числу в том же столбце слева.

Из только что доказанной теоремы следует, что если $y(x)$ есть многочлен степени n относительно x , то n -я разность будет

константой. Если бы мы хотели протабулировать многочлен степени n в нескольких равноотстоящих точках, то, в принципе, можно вычислить верхние числа в каждом столбце и построить всю остальную таблицу, пользуясь только сложением.

Для иллюстрации возьмем квадратный трехчлен

$$y(x) = 3x^2 - 6x + 9; \quad y(0) = 9$$

и вычислим его значения в точках от $x=0$ до $x=10$ с шагом $h=1$. Мы имеем

$$\begin{aligned} \Delta y(x) &= 6x - 3, & \Delta y(0) &= -3, \\ \Delta^2 y(x) &= 6, & \Delta^2 y(0) &= 6. \end{aligned}$$

Построим эту таблицу (таблица 1.4-3, отправные числа набраны жирным шрифтом). С другой стороны, можно было бы вычислить первые три значения и воспользоваться ими для отыскания разностей.

Таблица 1.4-3

$$y(x) = 3x^2 - 6x + 9$$

x	y	Δy	$\Delta^2 y$
0	9		
1	6	-3	
2	9	3	6
3	18	9	6
4	33	15	6
5	54	21	6
6	81	27	6
7	114	33	6
8	153	39	6
9	198	45	6
10	249	51	
Проверка: $y(10) = 300 - 60 + 9 = 249$			

В этом примере для вычисления очередного значения квадратного трехчлена потребовалось два сложения; вообще чтобы вычислить очередное значение многочлена n -й степени, требуется n сложений.

Большинство вычислений сейчас делается в системе плавающей запятой, но первоначальные исходные данные, которые часто снимаются непосредственно с автоматически регистрирующих приборов, бывают тогда в системе фиксированной запятой. При этих обстоятельствах указанный метод вычисления многочлена особенно полезен. Некоторое внимание следует, впрочем, уделить распространению ошибки вследствие использования приближенного значения разности в вычислениях.

Хотя таблицы 1.4-2 и 1.4-3 начинаются с нуля, часто полезно представлять таблицу распространяющейся неограниченно в обоих направлениях и, в частности,

представлять интересующее нас текущее место находящимся в нуле. Таблица 1.4-4 показывает, как единственная ошибка на одну единицу в функции $y(x) = 0$ распространяется в таблице разностей. Допустим, что ошибка была при $x=0$. Тогда $y(0) = 1$, а все другие значения $y(x) = 0$.

Таблица 1.4-4

Таблица распространения ошибки

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$	$\Delta^5 y$	$\Delta^6 y$
\vdots							
-3	0						
-2	0	0					
-1	0	0	0				
0	1	1	-2	-3	6	10	-20
1	0	-1	1	3	-4	-10	15
2	0	0	0	-1	1	5	-6
3	0	0	0	0	0	-1	1
4	0	0	0	0	0	0	\vdots
5	0	0	0	0	\vdots	\vdots	
6	0	\vdots	\vdots	\vdots			
\vdots	\vdots						

Числа в k -м столбце разностей равны $(-1)^r C(k, r)$. Чтобы доказать это, введем оператор сдвига E , который определяется правилом (принимая $h=1$)

$$E[f(x)] = f(x+1). \quad (1.4-1)$$

Ясно, что оператор линеен, так как

$$E[af(x) + bg(x)] = aE[f(x)] + bE[g(x)].$$

Повторное его применение дает

$$E^n[f(x)] = \underbrace{EEE \dots E}_{n \text{ раз}}[f(x)] = f(x+n).$$

Исходя из этого, мы определяем для $n=0$

$$E^0[f(x)] = f(x),$$

а для отрицательного показателя

$$E^{-n}[f(x)] = f(x-n).$$

Мы уже замечали, что иногда удобно представлять оператор отдельно от функции, на которую он действует. Если в тождестве отбросим настоящую функцию, то получим операторное равенство. Пример такого равенства:

$$\Delta = E - 1, \quad (1.4-2)$$

где 1 — тождественный оператор $1 \cdot f(x) = f(x)$. Правильность этого операторного равенства следует из простого замечания

$$\Delta y(x) = y(x+1) - y(x) = E[y(x)] - 1 \cdot y(x) = (E - 1)y(x).$$

Таким же образом следует, что

$$\Delta^k = (E - 1)^k = \sum_{r=0}^k (-1)^{k-r} C(k, r) E^r. \quad (1.4-3)$$

Для любой функции $y(x)$ операторное равенство для Δ^k дает формулу

$$\Delta^k y(n) = \sum_{r=0}^k (-1)^{k-r} C(k, r) y(n+r), \quad (1.4-4)$$

которую часто называют «разностной формулой Лагранжа» (основание для такого названия будет приведено позже). Из уравнения (1.4-4) можно найти отдельные разности, не вычисляя всей таблицы разностей.

Применим теперь это разложение к функции y целочисленного аргумента x , определенной двумя левыми столбцами таблицы 1.4-4:

$$y(x) = \begin{cases} 0, & x \neq 0, \\ 1, & x = 0. \end{cases}$$

Все члены суммы в правой части формулы (1.4-4) обратятся в нуль, кроме одного, который и даст соответствующий биномиальный коэффициент.

Формулу (1.4-4) можно использовать в обратную сторону, когда известно значение разности, а одного значения функции нет. В качестве характерного примера предположим известным, что шестые разности или в точности равны нулю, или настолько малы, что их можно считать приближенно равными нулю, а нам не хватает одного значения функции. Обозначим его y_0 , а смежные значения будут y_{-3} , y_{-2} , y_{-1} , y_1 , y_2 , y_3 . Тогда, используя (1.4-4) с $k=6$ и $n=-3$, имеем

$$y_{-3} - 6y_{-2} + 15y_{-1} - 20y_0 + 15y_1 - 6y_2 + y_3 = 0$$

или

$$y_0 = \frac{1}{20}[(y_{-3} + y_3) - 6(y_{-2} + y_2) + 15(y_{-1} + y_1)]. \quad (1.4-5)$$

Эта формула очень полезна, когда одно значение в гладкой таблице отсутствует из-за невнимательности. Однако подобной формулой надо пользоваться с осторожностью, так как отсутствующее значение может явиться следствием особенности в функции (например, деления на нуль). Заметим, что разности четного порядка использовать для этой цели много легче, чем разности нечетного порядка, так как для четных разностей имеется единственный максимальный член.

Упражнения

1.4-1. Вычислить $f(n) = n^2 - 79n + 1601$ для n от 0 до 10. (Замечание: $f(n)$ простое число для $n = 1, 2, \dots, 79$.)

1.4-2. Вычислить таблицу кубов от 10 до 20, используя таблицу разностей.

1.4-3. Предположим, что значение при $x = 0,6$ в таблице 1.4-2 отсутствует; считая, что четвертые разности равны нулю, вычислить $Si(0,6)$.

1.4-4. Вычислить приближенно $Si(1,1)$, используя (1.4-4).

§ 1.5. Факториалы

В анализе важную роль играет функция x^n . В ряде Тейлора

$$f(x) = \sum_{n=0}^{\infty} a_n x^n,$$

например, произвольная функция разложена по степеням x . Эта ведущая роль функции x^n в большой степени определяется тем, что она удовлетворяет соотношению

$$\frac{d}{dx} x^n = n x^{n-1} \quad (n \geq 1).$$

Естественная аналогия требует построения в исчислении разностей системы функций $g_n(x)$, которые удовлетворяют аналогичным соотношениям

$$\Delta g_n(x) = n g_{n-1}(x) \quad (n \geq 1). \quad (1.5-1)$$

Соотношению (1.5-1) удовлетворяют *факториальные многочлены* *), определенные равенствами

$$\begin{aligned} g_n(x) &= x^{(n)} = x(x-1)(x-2) \dots (x-n+1) \quad (n \geq 1), \\ g_0(x) &= x^0 = 1. \end{aligned} \quad (1.5-2)$$

Действительно,

$$\begin{aligned} \Delta g_n(x) &= g_n(x+1) - g_n(x) = \\ &= [(x+1) - (x-n+1)] x(x-1) \dots (x-n+2) = \\ &= n x(x-1) \dots [x - (n-1) + 1] = n x^{(n-1)} = n g_{n-1}(x). \end{aligned}$$

*) Не путайте с $n! = n(n-1)(n-2) \dots 2 \cdot 1$, которое является значением $x^{(n)}$ при $x = n$.

Заметим, что из $x^0 \equiv 1$ следует $0^{(0)} = 1$, соответствующее стандартному $0! = 1$, и что $x^{(n)}$ имеет ровно n сомножителей.

Пусть данная функция $f(x)$ может быть разложена по факториальным многочленам, аналогично разложению Тейлора по степеням x , т. е.

$$f(x) = b_0 + b_1 x + b_2 x^{(2)} + b_3 x^{(3)} + \dots = \sum_{k=0}^{\infty} b_k x^{(k)}.$$

Предположим также, что разностный оператор линеен для бесконечных сумм и что можно брать разности почленно. В этом разложении при $x=0$ получаем

$$f(0) = b_0.$$

Применяя оператор Δ к обеим частям равенства, получим

$$\Delta f(x) = \sum_{k=1}^{\infty} b_k k x^{(k-1)}.$$

Снова полагая $x=0$, имеем $\Delta f(0) = b_1$. Продолжая действовать таким образом, получим

$$\Delta^n f(x) = \sum_{k=n}^{\infty} b_k k(k-1) \dots (k-n+1) x^{(k-n)}$$

и, полагая $x=0$,

$$\Delta^n f(0) = n! b_n.$$

Мы получаем, следовательно, формальный аналог ряда Тейлора

$$f(x) = \sum_{k=0}^{\infty} \frac{\Delta^k f(0)}{k!} x^{(k)} = \sum_{k=0}^{\infty} C(x, k) \Delta^k f(0), \quad (1.5-3)$$

где $C(x, k)$ — биномиальные коэффициенты, которые могут быть записаны как $\frac{x^{(k)}}{k!}$.

Если ограничиваться многочленами, то ряд Тейлора будет содержать конечное число членов и вопрос о сходимости не возникнет. Аналогично разложение (1.5-3) для многочленов степени n будет конечным

$$f(x) = \sum_{k=0}^n \frac{\Delta^k f(0)}{k!} x^{(k)}, \quad (1.5-4)$$

так как по основной теореме (§ 1.3) $\Delta^{n+1} f(x) = 0$ и все разности более высокого порядка обратятся в нуль. Выражение (1.5-4) известно как *интерполяционная формула Ньютона*.

Определение факториала может быть обобщено тем же путем, каким пользуются для обобщения определения степени в алгебре. Используя определение (1.5-2), имеем

$$x^{(n)} = x^{(m)} (x - m)^{(n-m)}.$$

Если теперь формально положить $n=0$, то получим

$$x^{(0)} = 1 = x^{(m)} (x - m)^{(-m)}$$

или

$$(x - m)^{(-m)} = \frac{1}{x^{(m)}}.$$

Заменяя $x - m$ на y , приходим к формуле

$$y^{(-m)} = \frac{1}{(y+m)^{(m)}} = \frac{1}{(y+m)(y+m-1)\dots(y+1)}, \quad (1.5-5)$$

которая служит определением факториалов для отрицательных целочисленных показателей.

Заметим, что если $m \neq 0$, то

$$x^{(m)} x^{(-m)} \neq x^{(0)}.$$

Упражнения

1.5-1. Используя формулу Ньютона, найти многочлен, который принимает следующие значения (все $f(n)$ — простые числа при $n < 41$):

n	0	1	2	3	4	5
$f(n)$	41	43	47	53	61	71

О т в е т: $P(n) = n^2 + n + 41$.

1.5-2. Показать, что

$$3^{(-8)} = \frac{1}{120}, \quad 1^{(-m)} = \frac{1}{(m+1)!}.$$

§ 1.6. Деление многочленов

Деление многочленов иногда дается в элементарном курсе алгебры. Оно будет часто использоваться, а потому заслуживает внимательного рассмотрения. Мы покажем, как использовать деление многочленов для того, чтобы по коэффициентам многочлена найти коэффициенты его разложения по факториальным.

Если многочлен

$$f(x) = x^4 + 6x^3 - 7x + 8$$

Так как $R=f(a)$, то приведенный алгоритм можно рассматривать как способ вычисления значения многочлена. В этом случае схема принимает вид

$$y(x) = \{[(x+0)x+6]x-7\}x+8. \quad (1.6-2)$$

Заметим, с одной стороны, что для $n=6$ эта схема*) не является наиболее экономной по числу действий, если мы хотим многократно вычислять значения одного и того же многочлена шестой степени**) (например, в библиотечной программе). Так, схема

$$\begin{aligned} P_1 &= x(x+a_1), \\ P_2 &= (P_1+x+a_2)(P_1+a_3), \\ P &= (P_2+a_4)(P_1+a_5)+a_6 \end{aligned}$$

использует на одну операцию меньше. Коэффициенты a_i могут быть найдены через коэффициенты A_i многочлена

$$P(x) = x^6 + A_1x^5 + A_2x^4 + A_3x^3 + A_4x^2 + A_5x + A_6.$$

Вычисление коэффициентов a_i достаточно сложно, но его нужно сделать только один раз.

Приведем еще один пример использования деления многочленов. Если нужно вычислить значение многочлена $P(x)$ с действительными коэффициентами в комплексной точке

$$x_1 = a + ib \quad (i = \sqrt{-1}),$$

то можно образовать квадратный трехчлен

$$[x^2 - 2ax + (a^2 + b^2)]$$

и, разделив многочлен на этот действительный квадратный трехчлен, получить

$$P(x) = [x^2 - 2ax + (a^2 + b^2)] Q(x) + r_1x + r_2.$$

(Сокращенную формулу, подобную (1.6-1), можно написать и для квадратного трехчлена.) Положив $x=x_1$, найдем

$$P(x_1) = r_1x_1 + r_2.$$

В таком случае

$$P(x_1) + P(x_2) = r_1(x_1 + x_2) + 2r_2 = 2(r_1a + r_2);$$

часто бывает нужно вычислить не $P(x_1)$, а именно эту величину.

*) У нас ее принято называть схемой Горнера. (Прим. ред.)

**) Наблюдение принадлежит Моцкину. См. J. Todd, Motivation for Working in Numerical Analysis, Commun. Pure and Appl. Math., vol. 8, pp. 97—116 (1955). [Русский перевод: Тодд, Мотивы для работы в области численного анализа. Мат. просвещение, вып. 1 (1957), 76—86. См. также В. Я. Пан, Некоторые схемы для вычисления многочленов с вещественными коэффициентами, ДАН СССР 127, № 2 (1959), 266—269. (Прим. ред.)]

Вернемся теперь к задаче разложения данного многочлена в сумму по $x^{(n)}$. Чтобы сделать это, надо разделить последовательно многочлен на x , частное на $(x-1)$, следующее частное на $(x-2)$ и т. д., сделав n шагов. Так как деление на x тривиально, то для приведенного выше многочлена

$$P(x) = x^4 + 6x^3 - 7x + 8,$$

беря по очереди делители 0, 1, 2, 3, получаем

$$\begin{array}{r} 1 \overline{) 1 \ 0 \ 6 \ -7 \ +8} \\ \underline{1 \ 1 \ 7} \\ 2 \overline{) 1 \ 1 \ 7 } \\ \underline{2 \ 6} \\ 3 \overline{) 1 \ 3 \ 13} \\ \underline{3} \\ 1 \overline{) 6} \end{array}$$

или, в соответствии с (1.6-2),

$$P(x) = ([1(x-3) + 6](x-2) + 13)(x-1) + 0)x + 8, \quad (1.6-3)$$

$$P(x) = x^{(4)} + 6x^{(3)} + 13x^{(2)} + 0 \cdot x^{(1)} + 8. \quad (1.6-4)$$

Равенство (1.6-3) дает удобную форму для вычислительных работ, а (1.6-4) более отчетливо выделяет коэффициенты.

Упражнения

1.6-1. Представить x^3 в виде суммы факториальных многочленов.

1.6-2. Пусть $P(x) = x^3 - 5x^2 + 3x + 1$; вычислить $P(1+i) + P(1-i)$.
 Ответ: $x^{(3)} + 3x^{(2)} + x^{(1)}$.
 Ответ: 34.

§ 1.7*. Числа Стирлинга первого рода*)

Так как функции $x^{(n)}$ важны в исчислении разностей, а x^n — в дифференциальном исчислении, ясно, что могли бы быть полезны соотношения между этими функциями. Чтобы выразить $x^{(n)}$ через степени x , напомним

$$x^{(n)} = \sum_{k=0}^n S(n, k) x^k \quad (1.7-1)$$

и вычислим коэффициенты $S(n, k)$. Эти коэффициенты называются числами Стирлинга первого рода. Для $n=1$ находим

$$x^{(1)} = x = S(1, 0) + S(1, 1)x,$$

откуда $S(1, 0) = 0$; $S(1, 1) = 1$.

*) Параграфы 1.7 и 1.8, помеченные звездочкой, не будут использоваться в тексте нигде, кроме § 1.9.

Для $n=2$

$$x(x-1) = x^2 - x = S(2, 0) + S(2, 1)x + S(2, 2)x^2,$$

откуда

$$S(2, 0) = 0; \quad S(2, 1) = -1; \quad S(2, 2) = 1.$$

Вместо того чтобы действовать таким же образом дальше, получим общее рекуррентное соотношение, написав

$$x^{(n+1)} = (x-n)x^{(n)}.$$

Отсюда, используя дважды (1.7-1), получаем

$$\begin{aligned} \sum_{k=1}^{n+1} S(n+1, k) x^k &= (x-n) \sum_{k=1}^n S(n, k) x^k = \\ &= \sum_{k=1}^n [S(n, k-1) - nS(n, k)] x^k + S(n, n) x^{n+1} - nS(n, 0). \end{aligned} \quad (1.7-2)$$

Очевидно, что для всех n справедливо $S(n, n) = 1$, тогда как для $n > 0$ имеем $S(n, 0) = 0$. Приравняв коэффициенты в уравнении (1.7-2), получим рекуррентное соотношение, которое дает возможность находить числа Стирлинга последовательно:

$$S(n+1, k) = S(n, k-1) - nS(n, k) \quad (k=1, 2, \dots, n). \quad (1.7-3)$$

Для чисел Стирлинга нет простой формулы, как для биномиальных коэффициентов. Таблица 1.7-1 дает некоторые из этих чисел.

Т а б л и ц а 1.7-1

Числа Стирлинга первого рода $S(n, k)$

$\begin{matrix} k \\ n \end{matrix}$	1	2	3	4	5
1	1				
2	-1	1			
3	2	-3	1		
4	-6	11	-6	1	
5	24	-50	35	-10	1

Упражнение 1.7-1. Распространить таблицу 1.7-1 еще на одну строку, для $n=6$.

§ 1.8 *. Числа Стирлинга второго рода

Числа Стирлинга второго рода выражают x^n через факториальные многочлены

$$x^n = \sum_{k=1}^n \varphi(n, k) x^{(k)}. \quad (1.8-1)$$

Как и в предыдущем параграфе, вычислим сначала несколько значений, а затем найдем общее рекуррентное соотношение.

Для $n=1$ имеет место $x = \varphi(1, 0) + \varphi(1, 1)x$, откуда $\varphi(1, 0)=0$; $\varphi(1, 1)=1$. Для $n=2$ напомним $x^2 = \varphi(2, 0) + \varphi(2, 1)x + \varphi(2, 2)x(x-1)$, откуда $\varphi(2, 0)=0$; $\varphi(2, 1)=1$; $\varphi(2, 2)=1$.

Рекуррентное соотношение следует из равенства $x^{n+1} = x \cdot x^n$:

$$\begin{aligned} \sum_{k=1}^{n+1} \varphi(n+1, k) x^{(k)} &= x \sum_{k=1}^n \varphi(n, k) x^{(k)} = \sum_{k=1}^n \varphi(n, k) (x-k) x^{(k)} + \\ &+ \sum_{k=1}^n k \varphi(n, k) x^{(k)} = \sum_{k=1}^n \varphi(n, k) x^{(k+1)} + \sum_{k=1}^n k \varphi(n, k) x^{(k)} = \\ &= \sum_{k=1}^n [\varphi(n, k-1) + k \varphi(n, k)] x^{(k)} + \varphi(n, n) x^{(n+1)} - \varphi(n, 0). \end{aligned}$$

Ясно, что $\varphi(n, n)=1$ для любого n , тогда как для $n > 0$ имеем $\varphi(n, 0)=0$. Приравняв коэффициенты при подобных членах в факториалах (так как они линейно независимы), получим рекуррентное соотношение

$$\varphi(n+1, k) = \varphi(n, k-1) + k \varphi(n, k). \quad (1.8-2)$$

Из этого соотношения можно построить таблицу чисел Стирлинга второго рода (см. таблицу 1.8-1).

Таблица 1.8-1

Числа Стирлинга второго рода

$k \backslash n$	1	2	3	4	5
1	1				
2	1	1			
3	1	3	1		
4	1	7	6	1	
5	1	15	25	10	1

Упражнение 1.8-1. Распространить таблицу 1.8-1 еще на одну строку, для $n=0$.

§ 1.9. Пример

Приведем пример использования формулы Ньютона и чисел Стирлинга. Задача состоит в том, чтобы вычислить

$$g(y) = \frac{d}{dy} \int_0^y \frac{f(x)}{\sqrt{y-x}} dx, \quad 0 \leq y \leq 1,$$

если значения $f(x)$ заданы в точках $x=0; 0,1; 0,2; \dots; 1, 0$.

Очевидно, что при верхнем пределе, когда $x=y$, подынтегральное выражение становится неопределенным. Заметим также, что результат интегрирования должен быть продифференцирован по y ; но если дифференцировать по y под знаком интеграла, то полученный в результате интеграл будет расходиться. Таким образом, предложенная задача требует некоторого внимания.

Начнем с замечания, что если бы $f(x)$ было равно x^n , то нужно было бы вычислить простое выражение

$$g(y) = \frac{d}{dy} \int_0^y \frac{x^n dx}{\sqrt{y-x}}.$$

Здесь естественно положить

$$x = y \sin^2 \theta, \quad dx = 2y \sin \theta \cos \theta d\theta,$$

и мы получим, что для $f(x) = x^n$ искомая функция равна

$$g(y) = \frac{d}{dy} \frac{2y^{n+1}}{\sqrt{y}} \int_0^{\pi/2} \sin^{2n+1} \theta d\theta = \frac{(2n+1)y^n}{\sqrt{y}} W_{2n+1},$$

где

$$W_{2n+1} = \int_0^{\pi/2} \sin^{2n+1} \theta d\theta = \frac{2 \cdot 4 \dots (2n)}{1 \cdot 3 \cdot 5 \dots (2n+1)} = \frac{2n}{2n+1} W_{2n-1} \quad W_1 = 1.$$

Далее, заметим, что если

$$f(x) = \sum_{n=0}^{10} a_n x^n,$$

то

$$g(y) = \frac{1}{\sqrt{y}} \sum_{n=0}^{10} (2n+1) W_{2n+1} a_n y^n.$$

Если теперь допустить, что истинная функция $f(x)$ может быть аппроксимирована достаточно точно многочленом десятой степени по

данным узловым точкам, то остается только найти a_n . Один из способов нахождения a_n — использование формулы Ньютона (1.5-4) с шагом по $x=0,1$:

$$f(x) = f(0) + 10x\Delta f(0) + \frac{10x(10x-1)}{2!} \Delta^2 f(0) + \dots + (10x)^{(10)} \Delta^{10} f(0)$$

(в разностях принимается $h=0,1$), чтобы затем через числа Стирлинга первого рода (уравнение (1.7-1)) получить коэффициенты a_n при степенях x . Как будет показано позже, использование десятых разностей может оказаться опасным; но в задаче, из которой родился этот пример, где данные были грубыми и ожидаемый результат не намного лучше, он выглядел удовлетворительным. Об этом можно судить по следующим фактам: (1) график аппроксимирующего многочлена выглядел разумно; (2) результаты были совместимыми с другими частями физической теории, которая изучалась, и (3) результаты стимулировали дальнейшую работу.

§ 1.10. Альтернативные замечания

Выбор определения разностного оператора (равенство (1.2-1))

$$\Delta f(x) = f(x+h) - f(x)$$

был произвольным. Можно было бы вместо этого выбрать обратные разности

$$\nabla f(x) = f(x) - f(x-h).$$

Что произошло бы, если бы мы это сделали? В частности, возрастающие факториальные многочлены *)

$$^{(n)}x = x(x+h)(x+2h) \dots [x+(n-1)h]$$

дали бы, аналогично (1.5-1),

$$\nabla^{(n)}x = n[^{(n-1)}x].$$

Отсюда можно получить обратную формулу Ньютона, соответствующую (1.5-4). Аналогично возникли бы новые числа Стирлинга первого и второго рода $\sum(n, k)$ и $\sigma(n, k)$. Они связаны со старыми числами Стирлинга равенствами

$$\sum(n, k) = (-1)^{n+k} S(n, k) = |S(n, k)|, \quad \sigma(n, k) = (-1)^{n+k} \varphi(n, k).$$

Таким образом, теория, совершенно аналогичная теории, которая была развита выше, могла бы основываться на обратных разностях. Формулы, которые получаются из такой теории, полезны специалисту, но, вероятно, являются роскошью для потребителя.

*) В отличие от введенных в § 1.5 они обозначаются $^{(n)}x$. (Иприм. ред.)

Центральные разности и средние разности приводят еще к двум способам записи, которые иногда применяются, но они также являются просто способами записи и дают небольшое удобство лишь в некоторых случаях; следовательно, они, по-видимому, не стоят беспокойства потребителя.

§ 1.11. Общие замечания и справки

Мы дали лишь краткое введение в исчисление конечных разностей, хотя на эту тему написаны целые книги. Одной из лучших является книга Жордана [19]*).

Исчисление конечных разностей может быть применено для получения многих интересных результатов, таких как

$$\sum_{k=0}^n S(n, k) \varphi(k, m) = \delta(n, m),$$

где $\delta(n, m)$ — «символ Кронекера», равный нулю, если $m \neq n$, и единице, если $m = n$. Справедливость формулы следует из простого замечания, что если разложить факториал по степеням x , а затем, наоборот, степени x по факториалам, то все коэффициенты, кроме коэффициента при начальном факториале, обратятся в нуль. Такие результаты редко применяются в практических вычислениях, но интересны с точки зрения математики.

Многие книги используют операторные методы. Мы используем лишь операторы Δ и E в нескольких простых конечных случаях и тщательно избегаем применения подозрительных бесконечных процессов с символическими операторами. Операторные методы часто являются многообещающими, но их применение предполагает хорошее знакомство с ними, которого потребитель может и не иметь. Кроме того, автор полагает, что формальные манипуляции часто скрывают смысл формул и, следовательно, затрудняют понимание исследуемой задачи.

ГЛАВА 2

ПОГРЕШНОСТИ ОКРУГЛЕНИЯ

§ 2.1. Введение

Уже при ручных вычислениях сравнительно рано приходится сталкиваться с эффектами округления, возникающими вследствие того, что в процессе вычислений сохраняется лишь конечное число десятичных знаков. Впрочем, для ручных вычислений эта проблема менее

*.) Числа в квадратных скобках относятся к библиографии в конце книги.

важна, нежели для машинных. Это объясняется несколькими причинами.

Во-первых, объем работ, которые могут быть выполнены вручную, весьма ограничен по сравнению с тем, что делается на современных быстродействующих вычислительных машинах. Во-вторых, в процессе ручного счета человек может непосредственно наблюдать многие из эффектов округления и предпринять требуемые меры в случае, когда нужно предостеречься от ошибки. В-третьих, ручные вычисления обычно производятся способом, который может быть назван «вычисление с переменной длиной числа, с квазификсированной-квазиплавающей запятой», в котором длина числа регулируется так, чтобы избежать грубых ошибок округления; машинное же вычисление ведется обычно с плавающей запятой и с фиксированной длиной числа. В-четвертых, при ручных вычислениях обычно нетрудно оценить максимальную величину ошибки, которая может возникнуть вследствие округления. В машинных вычислениях получение такой оценки очень дорого, из-за чего приходится прибегать к статистическим оценкам.

Таким образом, в универсальных вычислительных машинах часто бывает неизбежен статистический подход к оценкам погрешности: лучше действовать с известным риском, чем не действовать вовсе. По этим причинам ранее разработанные теории оценки эффектов округления, как правило, непригодны для современных машинных вычислений. Действительно удовлетворительных теорий, пригодных для этой цели, в настоящее время нет. Мы ограничимся лишь кратким обсуждением некоторых фрагментов таких теорий, из которых многие находятся в зачаточном состоянии.

Из вычислительной практики возникло несколько широко применяемых методов использования вычислительной машины для установления наличия ошибок, а также и некоторые методы использования вычислительной машины для оценки величины ошибки. Последнее особенно необходимо, так как прежде, чем писать программу для машины, важно иметь какие-то соображения об ожидаемой точности.

§ 2.2. Область ответа

По-видимому, самым простым и успешным подходом к проблеме округления является определение *области ответа*. Каждая величина изображается здесь на самом деле двумя числами: максимальным и минимальным значениями, которые она может принимать. В некотором смысле каждое число заменяется областью, в которой лежит точное значение величины; этим и объясняется название. При выполнении действий над величинами новая область вычисляется соответствующим образом из данных областей, используя подходящие округления. Поэтому на каждой стадии вычислений существуют надежные границы, в которых лежит верный ответ.

Применение этого приема можно осуществить переделкой машины или программой, заменяющей такую переделку. В обоих случаях это более чем удваивает работу и требует вдвое больше памяти, чем при обычных вычислениях.

В задачах средней сложности такой подход может быть успешно применен, и получением областей определения можно фактически воспользоваться. С другой стороны, в сравнительно больших задачах границы области ответа часто так далеки одна от другой, что ответ является почти никчемным.

Несмотря на то, что верный ответ часто действительно находится около середины полученной области, надеяться на это опасно. В самом деле, предположим, что последняя выполненная операция есть умножение и что оба сомножителя имели область от 0 до 1. Если считать, что ответы находятся около середины области, то оба сомножителя должны иметь верные значения около 0,5. Но окончательный результат имел бы тогда ту же самую область, а его верное значение находилось бы около 0,25, что лежит далеко от середины области и потому противоречит сделанному допущению. Таким образом, ощущение, что верное значение должно быть около середины области, не всегда оправдано. С несомненностью мы добились лишь знания области, но у нас нет никаких сведений относительно того, в какой части области лежит верный ответ.

Упражнение 2.2-1. Составить блок-схему вычисления области в случае произведения двух чисел, когда известны области этих чисел.

§ 2.3. Двойная точность

Другим распространенным методом оценки является решение задачи *с обычной и двойной точностью*. Принято верить, что совпадающие в двух ответах разряды верны. Выполнив дважды несколько типичных вычислений, мы можем вести остальной счет с обычной точностью, полагая, что и в них верно то же самое количество разрядов.

Опасности этого метода очевидны. Тем не менее он широко применяется и с удовлетворительными результатами. Он несколько напоминает «область ответа» хотя бы тем, что требует такого же увеличения памяти и объема вычислений. Однако этот метод дает обычно более точные ответы, чем метод «области ответа».

§ 2.4. Счет со значащими разрядами

Основным соображением, которое лежит в основе *счета со значащими разрядами*, является то, что большие потери точности, как правило, происходят при вычитании двух близких чисел. При этом образуется несколько нулей в начале числа, которые потом удаляются

при нормализации в операциях с плавающей запятой. Если заблокировать последний сдвиг, сохранив стоящие впереди нули, то результат будет более или менее верно показывать число значащих разрядов.

Возникло два варианта такого метода. В одном из них сохраняются ведущие нули и выдаются числа, которые содержат лишь значащие разряды. В другом — сдвиг фактически делается, но число нулей сохраняется в виде индекса, аналогичного показателю в системе с плавающей запятой.

Оба варианта нетрудно воспроизвести соответствующими программами. Опыт работы с обоими вариантами этого метода мог бы более подробно осветить их роль в вычислениях.

§ 2.5. Статистический подход

Статистический подход базируется на заманчивой мысли, что округление есть случайный процесс, и, следовательно, можно попытаться построить модель округления, основываясь на теории вероятностей. Этот подход начинается со следующего парадокса: если повторить вычисление, то результат окажется в точности тем же самым (в предположении, что машина работает правильно) и, следовательно, результат не является случайной переменной.

Ситуация несколько напоминает бросание монетки. При каждом бросании мы полагаем, что если достаточно много знать о силах и распределении веса в монетке, то можно подсчитать, упадет ли она на герб или на решку. Мы предпочитаем, однако, не прибегать к такой большой работе и представить себе модель, в которой герб или решка есть случайная переменная. Таким же образом, хотя и предполагается, что можно высчитать результат, мы предпочитаем рассматривать его как случайную переменную. Полученное число рассматривается как верный ответ плюс эффект случайного округления.

Для того чтобы формально применить статистику, нужно создать (на самом деле или мысленно) множество ответов. Хотя существует много путей создания множества, примем временно следующую умозрительную модель. Представим, что вычисление может быть сделано с бесконечной точностью, и мы хотим присоединить погрешности округления после каждого арифметического шага — погрешности, аналогичные помехам в теории информации, где они обычно рассматриваются как присоединенные к ожидаемому сигналу.

Наша задача теперь — найти распределение множества или по крайней мере оценить это распределение. Произведем на машине одно вычисление. Если полученный ответ характерен для множества ответов построенной нами модели, то мы сможем сделать обоснованные выводы для дальнейшего счета. Если нет, то не сможем. Однако в любом случае полученный нами результат обладает свойствами, отличными от свойств большей части множества ответов модели.

Действительно, допустимые для машины погрешности округления обладают тем специфическим свойством, что в результате округления получается число с нулями, начиная с некоторого разряда. Поэтому, вероятно, наш ответ будет обладать также и некоторыми другими частными свойствами.

Разные теории округления отличаются главным образом тем, как они пытаются найти распределение множества, которое мы представили. Ясно, что мы не сможем в действительности составить множество и должны будем оценить его по тем немногим сведениям, которые сможем получить.

§ 2.6. Случайное округление

Одним из очевидных методов получения образцов требуемого множества является следующий: построить *случайное округление* либо воспроизводящей программой, либо действительными изменениями машины, и повторить вычисления несколько раз. Можно надеяться из распределения полученных ответов получить правильные оценки истинного распределения множества.

Основное возражение против этого метода: требование большого количества времени и денег. Действительно, нужно много раз повторить счет, чтобы получить достаточно надежную оценку отдельных флуктуаций и, таким образом, иметь возможность судить, насколько можно быть уверенным в ответе. Даже чтобы оценить среднее значение с достаточной точностью, требуется довольно много случаев.

§ 2.7. Переменная точность

В последнее время появилось предложение повторять решение задачи несколько раз, причем при каждом проходе держать на один разряд меньше, чем в предыдущем. Исходя из полученных решений формулируется оценка точности наиболее точного решения.

Интуитивно кажется очевидным, что такое же число проходов со случайным округлением, как в § 2.6, дало бы более надежную оценку (так как в этом случае ответы были бы более тесно связаны с искомым результатом). Однако этот вопрос не исследован в достаточной степени ни теоретически, ни экспериментально.

§ 2.8. Оценка шума в таблице

Предлагаемый метод применим в том случае, когда составлена таблица чисел и каждое число имеет ошибку округления, не зависящую от остальных. Это — та частая в практической работе ситуация, когда нужно сосчитать много вариантов задачи для равноотстоящих узлов независимого переменного. Таким образом, мы пытаемся

ответить на следующий вопрос: пусть дана таблица ответов вычисления $f(x)$ для группы равноотстоящих x ; что является предполагаемым «шумом округления» в этих ответах?

Заметим, что мы никоим образом не пытаемся найти никаких систематических ошибок. Мы хотим лишь, получив результат и добавив еще немного вычислений, оценить шум в ответах.

Таблица 1.4-1 показывает, как единственная ошибка распространяется в таблице разностей. Используя линейное свойство оператора Δ , можно представить таблицу разностей затабулированной функции как сумму таблицы разностей истинной функции и таблицы разностей шума. Эта последняя в свою очередь может быть представлена как сумма таблиц разностей элементарных шумов, причем r -я таблица соответствует шуму ε_r на r -м месте в первом столбце таблицы шума.

Суммарный шум в k -й разности, в соответствии с (1.4-1), есть

$$\Delta^k \varepsilon_n = \sum_{r=0}^k (-1)^{k-r} C(k, r) \varepsilon_{n+r}. \quad (2.8-1)$$

Его границы могут быть легко найдены из ограничения на отдельный шум в таблице. Если ε_r ограничено величиной ε , т.е. $|\varepsilon_r| \leq \varepsilon$, то

$$|\Delta^k \varepsilon_n| \leq \sum_{r=0}^k C(k, r) \varepsilon = 2^k \varepsilon. \quad (2.8-2)$$

Для $k=10$, например, это представляет тысячекратное увеличение шума.

Поскольку границы шума в k -й разности так пессимистичны, имеет смысл возвратиться к статистическому подходу. Естественно предположить, что отрицательная и положительная ошибки ε_n равновероятны, и поэтому среднее значение $M(\varepsilon)$ множества ε_n есть нуль. Нас интересует дисперсия $D(\Delta^k \varepsilon_n)$ распределения k -й разности. Пусть дисперсия $D(\varepsilon_n)$ шума округления ε_n есть σ^2 , так что

$$\sigma^2 = D(\varepsilon_n) = M_2(\varepsilon_n^2) - M^2(\varepsilon_n) = M(\varepsilon_n^2),$$

где среднее значение берется, конечно, по множеству. Для дисперсии k -й разности, используя (2.8-1) и тот факт, что среднее значение ε_n есть нуль, получим

$$D(\Delta^k \varepsilon_n) = M \left[\sum_{r=0}^k (-1)^{k-r} C(k, r) \varepsilon_{n+r} \right]^2.$$

Выполнив возведение в квадрат, находим

$$\begin{aligned} M \left[\sum_{r,s} (-1)^r (-1)^s C(k, r) C(k, s) \varepsilon_{n+r} \varepsilon_{n+s} \right] &= \\ &= M \left[\sum_r C^2(k, r) \varepsilon_{n+r}^2 + 2 \sum_{r < s} (-1)^{r+s} C(k, r) C(k, s) \varepsilon_{n+r} \varepsilon_{n+s} \right]. \end{aligned}$$

Для упрощения этого выражения используем следующие обстоятельства. По предположению шумы округления ϵ_n независимы и $M(\epsilon_n) = 0$. Отсюда следует, что $M(\epsilon_{n+r} \cdot \epsilon_{n+s}) = 0$ при $r \neq s$. Кроме того, известно, что

$$\sum_{r=0}^k C^2(k, r) = C(2k, k)$$

(упражнение 2.8-2). Таким образом, получаем

$$\begin{aligned} M \left[\sum_r C^2(k, r) \epsilon_r^2 \right] &= \\ &= \sum_r C^2(k, r) M(\epsilon_r^2) = \sigma^2 C(2k, k) = \frac{(2k)!}{(k!)^2} \sigma^2. \quad (2.8-3) \end{aligned}$$

Таблица 2.8-1 дает представление о росте коэффициента усиления шума, который есть просто $\sqrt{C(2k, k)}$.

Т а б л и ц а 2.8-1

Коэффициенты усиления округления

Порядок разности	Дисперсия $C(2k, k)$	Среднеквадратичное значение шума $\sqrt{C(2k, k)}$	Максимум шума 2^k	Порядок разности	Дисперсия $C(2k, k)$	Среднеквадратичное значение шума $\sqrt{C(2k, k)}$	Максимум шума 2^k
1	2	1,414	2	4	70	8,367	16
2	6	2,449	4	5	252	15,875	32
3	20	4,472	8	6	924	30,397	64

Теоретический результат, который здесь получен, требует усреднения величины из множества ϵ_n . На практике это сделать невозможно и мы прибегаем к обычному в такой ситуации приему: заменяем среднее значение множества ϵ_n средним значением функции (стандартный прием, использующийся в эргодической теории).

Можно было бы думать, что среднее значение k -х разностей в таблице 1.4-1 дает неплохую оценку среднего значения множества (в предположении, что применяется эргодический принцип), но, к сожалению, k -е разности сильно коррелированы.

Действительно, коэффициент корреляции для смежных значений равен

$$\begin{aligned} \sum_r \frac{(-1)^r C(k, r)}{\sqrt{C(2k, k)}} \frac{(-1)^{r-1} C(k, r-1)}{\sqrt{C(2k, k)}} &= - \sum_r \frac{C(k, r) C(k, r-1)}{C(2k, k)} = \\ &= - \frac{C(2k, k+1)}{C(2k, k)} = - \frac{k}{k+1} \quad (2.8-4) \end{aligned}$$

(см. упражнение 2.8-3). Когда $k \rightarrow \infty$, коэффициент корреляции стремится к -1 и показывает, что числа в k -м столбце случайны, но не независимы, как часто полагают; они имеют устойчивую тенденцию чередовать знаки $+$, $-$, $+$, $-$, ...

Сильная отрицательная корреляция у последовательных k -х разностей дает возможность обнаруживать отдельные ошибки в таблице. Этот метод лучше всего понимается на примере. Рассмотрим таблицу 2.8-2, где, как обычно, пишутся лишь последние знаки разностей. Поведение четвертых разностей, наибольшие значения которых сгруппированы около 35° , заставляет предполагать ошибку в функции для 35° . Поэтому мы предполагаем, что четвертые разности постоянны и что при 35° есть ошибка. Так как структура четвертых разностей известна (уравнение (1.4-3)),

$$1, -4, 6, -4, 1,$$

то можно попытаться приближенно удовлетворить следующим уравнениям:

$$41 = -4\varepsilon + C, \quad -59 = 6\varepsilon + C, \quad 37 = -4\varepsilon + C.$$

Решая их, получаем приблизительно $\varepsilon = -10$, так что правильное значение функции в 35° должно быть 6,421. Подставив 6,421 вместо 6,411, получим таблицу разностей:

				-4
			8	
		-50		1
	-468		9	
		-41		1
6,421	-509		10	
		-31		-3
			7	
				0

Продемонстрировав сильную корреляцию, которая существует между последовательными числами в столбце k -х разностей, и проиллюстрировав, как этим можно воспользоваться, вернемся к нашей главной задаче оценки шума в таблице. Предварительно предположим, что дисперсия σ^2 шума в таблице значений функции известна; мы хотим оценить дисперсию шума в k -м столбце. Если истинные значения функции в нашей таблице были значениями многочлена, его разности, начиная с некоторого порядка, должны были бы обратиться в нуль, а то, что осталось бы в таблице разностей, было бы следствием шума. Затруднения состоят в том, что функция не обязательно является многочленом; более того, даже если эта функция — многочлен, то, не зная его степени, мы не знаем, какого порядка разности нужно взять.

Т а б л и ц а 2.8-2

Гра- дусы	$C_n(z)$	Δ	Δ^2	Δ^3	Δ^4	Теорети- ческие четвертые разности вследствие ошибки ε
0	8,346					
5	8,302	— 44	—87			
10	8,171	—131	—83	4	3	
15	7,957	—214	—76	7	— 1	
20	7,667	—290	—70	6	6	
25	7,307	—360	—58	12	—14	$+s$
30	6,889	—418	—60	— 2	41	$-4s$
35	6,411	—478	—21	39	—59	$+6s$
40	5,912	—499	—41	—20	37	$-4s$
45	5,372	—540	—24	17	—10	$+s$
50	4,808	—564	—17	7	— 2	
55	4,227	—581	—12	5	0	
60	3,634	—593	— 7	5	— 2	
65	3,034	—600	— 4	3	— 2	
70	2,430	—604	— 3	1	2	
75	1,823	—607	— 0	3	— 4	
80	1,216	—607	— 1	— 1	2	
85	0,608	—608	0	1		
90		—608				

На практике в благоприятных случаях разности имеют тенденцию сначала уменьшаться по величине, а затем увеличиваться, одновременно испытывая сильные колебания из-за отрицательной корреляции. В этом случае обычно считают первый столбец разностей, следующий за минимальным, существующим вследствие шума, хотя, насколько известно автору, соответствующие исследования по этому вопросу не

проводились. Таким образом, среднее значение квадратов разностей в этом столбце мы принимаем за оценку квадратов разностей, усредненных по множеству. Разделив это среднее значение на $C(2k, k)$, получаем оценку квадрата уровня шума в исходной таблице.

Таблица 2.8-3

Гамма-функция $\Gamma(x)$

x	$\Gamma(x)$	Δ	Δ^2	Δ^3	Δ^4
1,0	1,000				
1,1	0,951	-49	+16		
1,2	0,918	-33	+12	-4	3
1,3	0,897	-21	+11	-1	-1
1,4	0,887	-10	+9	-2	+2
1,5	0,886	-1	+9	0	-2
1,6	0,894	+8	+7	-2	+2
1,7	0,909	+15	+7	0	+2
1,8	0,931	+22	+9	+2	
1,9	0,962	+31			

В качестве примера рассмотрим таблицу округленных значений $\Gamma(x)$, приведенных в таблице 2.8-3. Среднее значение $[\Delta^4 \Gamma(x)]^2 = (9 + 1 + 4 + 4 + 4 + 4) : 6 = \frac{13}{3}$. Разделим его на $C(8, 4) = 70$:

$$\frac{13}{3 \cdot 70} = 0,062 = \sigma_1^2 \text{ — оценка шума в таблице.}$$

Теоретическая оценка основана на равномерном распределении случайного округления

$$\sigma_2^2 = \int_{-1/2}^{1/2} x^2 dx = \frac{1}{12} = 0,0833. \quad (2.8-5)$$

Ввиду того, что таблица, взятая в этом примере, невелика, близость полученных оценок

$$\sigma_1 = 0,25, \quad \sigma_2 = 0,29$$

можно считать хорошим подтверждением теории. Точное вычисление ошибки округления дает

$$\sigma_3 = 0,32.$$

Упражнения

2.8-1. Раскрыв обе части равенства $(1+t)^{a+b} = (1+t)^a (1+t)^b$ и приравняв коэффициенты при одинаковых степенях t , получить

$$C(a+b, r) = \sum_{s=0}^a C(a, s) C(b, r-s).$$

2.8-2. В упражнении 2.8-1, положив $a=b=r$, получить

$$C(2r, r) = \sum_{s=0}^r C^2(r, s).$$

2.8-3. В упражнении 2.8-1, положив $a=b=n$, $r=n+1$, получить

$$C(2n, n+1) = \sum_{s=0}^n C(n, s) C(n, s-1).$$

§ 2.9. Теория «младшего значащего разряда»

Рассмотрим одну из двух новых теорий, цель которых показать, как развиваются эффекты округления. Они могут быть названы «теории округления в малом». Откажемся здесь от статистической модели § 2.5 и построим наше множество, полагая числа, над которыми производятся арифметические операции, варьирующимися. Таким образом можно будет изучать, как распространяются ошибки округления в типичном небольшом куске вычислений.

Когда вычисленные числа округляются, кажется разумным предположить, что совершаемая ошибка (в действительности изменение вследствие округления) равномерно распределена между $-1/2$ и $1/2$ последнего значащего разряда. Другими словами, любое значение ошибки столь же возможно, как любое другое в этом интервале. На вероятностном языке вероятность $P\{x_1 \leq \varepsilon \leq x_2\} = x_2 - x_1$ ($x_1 \leq x_2$ в $(-1/2, 1/2)$) (рис. 2.9-1).

В этой модели допускается непрерывное распределение и игнорируется тот очевидный факт, что действительное распределение машинного округления должно быть дискретным, так как могут встречаться лишь некоторые округления (например, потому что машина оперирует с числами конечной длины).

Посмотрим теперь, что происходит, когда складывается несколько чисел. Будем предполагать, что все числа в плавающем виде имеют один и тот же порядок и что он не меняется ни в одном из сложений и вычитаний.

Для начала рассмотрим сложение двух чисел. Очевидно, ошибка суммы лежит в интервале от -1 до 1 последнего значащего разряда.

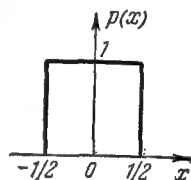


Рис. 2.9-1. Плотность распределения ошибки округления для одного числа.

Пусть ε_1 — ошибка первого слагаемого, ε_2 — ошибка второго, ε — ошибка суммы. Рассмотрим вероятность того, что $\varepsilon = -a$ ($a \geq 0$). Выбор ε_1 ограничен интервалом от $-1/2$ до $1/2 - a$, а ε_2 определяется исключительно выбором ε_1 . Таким образом, плотность вероятности есть

$$P'(a) = p(a) = \int_{-1/2}^{1/2-a} d\varepsilon_1 = 1 - a.$$

Аналогично поступаем применительно к положительным ошибкам. Таким образом, мы получаем треугольное распределение, изображенное на рис. 2.9-2.

Вместо того чтобы рассматривать суммы трех чисел, затем четырех, пяти и т. д., обратимся к центральной предельной теореме,

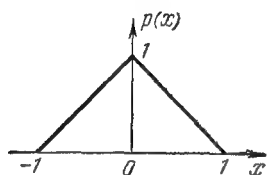


Рис. 2.9-2. Плотность распределения ошибки округления для суммы чисел.

которая, по крайней мере интуитивно, широко известна. В нашем случае, когда среднее значение ошибки равно нулю, эта теорема гласит, что если число округленных чисел, которые складываются, увеличивается, то функция распределения для ошибки суммы стремится к нормальному распределению

$$P\{x \leq \varepsilon \leq x + dx\} = \frac{1}{\sigma \sqrt{2\pi}} e^{-1/2 (x/\sigma)^2} dx$$

с некоторым соответствующим σ , которое будет определено позже. Фактически эта сходимость столь быстрая, что часто сумма 10 или 12 случайных чисел, выбранных из равномерного распределения, используется как хорошее приближение к единственному числу, выбранному из нормального распределения (см. § 32.4). Мы уже вычислили

$$D(\varepsilon) = \int_{-1/2}^{1/2} \varepsilon^2 d\varepsilon = \frac{1}{12}. \quad (2.9-1)$$

Для суммы независимых ошибок округления $\varepsilon = \varepsilon_1 + \varepsilon_2 + \dots + \varepsilon_n$ имеем

$$D(\varepsilon) = D\left(\sum_i \varepsilon_i\right) = \iint \dots \int \left(\sum_i \varepsilon_i\right)^2 d\varepsilon_1 d\varepsilon_2 \dots d\varepsilon_n.$$

Используя равенства

$$\int_{-1/2}^{1/2} d\varepsilon_i = 1 \quad \text{и} \quad \int_{-1/2}^{1/2} \varepsilon_i d\varepsilon_i = 0,$$

получим

$$D(\varepsilon) = \sum_{i,j} \iint \varepsilon_i \varepsilon_j d\varepsilon_i d\varepsilon_j = \sum_i \int \varepsilon_i^2 d\varepsilon_i = \frac{n}{12} = \sigma^2.$$

Таким образом, приравнивая дисперсию двух распределений, получаем $\sigma^2 = n/12$. Если n достаточно велико, нормальное распределение является хорошей аппроксимацией. Для нормального распределения существует симметричный интервал, в котором с вероятностью 50% лежит ошибка:

$$\text{от } -0,6745\sigma \text{ до } +0,6745\sigma.$$

Таким образом, мы видим, что длина половины интервала растет как \sqrt{n} , тогда как полный интервал растет как n . Это — одна из причин того, что метод области ответа, который, естественно, дает полную область, так пессимистичен по сравнению со статистическими оценками.

Эту модель не следует принимать всерьез, несмотря на сложное математическое изложение, так как гипотеза, что во время сложений не происходит сдвигов при выравнивании порядков и нормализации, слишком нереалистична. Каждый раз, когда происходит сдвиг, имеется тенденция возвратиться к равномерному распределению. Таким образом, при сложении и вычитании на последние значащие разряды действуют как бы две противоположные силы: одна толкает в направлении нормального распределения, тогда как другая — к равномерному.

Упражнение 2.9-1. Используя распределение ошибки округления суммы двух чисел (рис. 2.9-2), получить распределение ошибки суммы четырех чисел.

§ 2.10. Теория «старшего значащего разряда»

Чтобы понять распространение ошибок округления при операциях умножения и деления, необходимо исследовать распределение старших значащих разрядов. Пусть x_1 и x_2 — два числа с ошибками округления ϵ_1 и ϵ_2 . Если эти числа перемножить, то получим

$$(x_1 + \epsilon_1)(x_2 + \epsilon_2) = x_1x_2 + x_1\epsilon_2 + x_2\epsilon_1 + \epsilon_1\epsilon_2.$$

Если шум вследствие округления достаточно высок, дополнительные ошибки округления, которые возникают в процессе умножения, малы по сравнению с шумом округления в произведениях $x_1\epsilon_2$ и $x_2\epsilon_1$. Таким образом, старшие разряды в x_1 и x_2 влияют на распространение ошибок округления через умножение и, аналогичным образом, через деление.

Мы хотим показать, что разные старшие разряды встречаются неодинаково часто. Так как это обычно воспринимается с удивлением, мы, во-первых, обратимся к некоторым экспериментальным данным. Сто физических констант, имеющих физическую размерность, были выбраны случайным образом (не допускались безразмерные

отношения или чистые числа) из «Справочника по физике и химии» [15]. Наблюдаемое распределение старшего разряда показано в таблице 2.10-1.

В третьем столбце стоят теоретические значения, вычисленные по формуле $100[\ln(n+1) - \ln n]$, описывающей логарифмическое распределение. Маленькие разности показывают, что в теории, утверждающей, что старшие разряды распределены логарифмически, есть какой-то смысл.

Из физических соображений можно привести много разных аргументов, чтобы подтвердить этот результат. Прежде всего рассмотрим распределение старших разрядов во всех физических константах.

Таблица 2.10-1

Распределение старших разрядов случайно выбранных физических констант

Ведущий разряд	Записанное число	Теоретическое число	Разность
1	34	30	+4
2	12	18	-6
3	13	12	+1
4	15	10	+5
5	7	8	-1
6	3	7	-4
7	4	6	-2
8	4	5	-1
9	8	4	+4
Итого	100	100	0

Эти константы рассеяны в большом интервале чисел. Посмотрим теперь, к чему приведет изменение масштаба. Физические константы, естественно, тоже изменятся, но трудно поверить, что вид распределения старших разрядов изменится очень сильно. Если же изменений не происходит вообще, то можно показать, что распределение логарифмическое.

Другой интуитивный аргумент за то, что распределение не будет равномерным, заключается в следующем. Рассмотрим одноразрядную вычислительную машину с плавающей запятой и предположим, что распределение старших

разрядов равномерно. Если исследовать все суммы или произведения, которые могут быть получены, то обнаружится сильное смещение в сторону чисел с меньшими старшими разрядами.

Это рассуждение может быть проведено более строго. Будем считать, что мантисса z нормализованного числа принимает значения от 1 до 10, именно

$$1 \leq z < 10.$$

Предположим, что z распределено в этом интервале равномерно с плотностью $1/9$. Можно получить следующие результаты.

1. Мантисса произведения двух чисел, выбранных независимо из упомянутого распределения, имеет плотность распределения

$$\frac{10 \ln 10 - 9 \ln z}{81}.$$

2. Частное имеет плотность распределения

$$\frac{1}{18} \left(1 + \frac{10}{z^2} \right).$$

3. Длинная последовательность независимых умножений или умножений и делений с любым (разумным) первоначальным распределением всегда приводит к распределению, плотность которого быстро стремится к предельной функции

$$\frac{1}{z \ln 10}.$$

Кривые для этих трех распределений показаны на рис. 2.10-1 вместе с первоначальным равномерным распределением, из которого выбирались числа.

Докажем первое утверждение. Посмотрим, как в произведении может получиться число с мантиссой z . Выберем произвольно первый множитель. Плотность вероятности выбранного значения есть $\frac{1}{9}$. Теперь, если $x_1 \leq z$, то второй множитель определяется

однозначно и равен $\frac{z}{x_1}$, тогда как если $x_1 \geq z$, то второй множитель есть $\frac{10z}{x_1}$ и в любой из этих точек плотность вероятности равна $\frac{1}{9}$. Таким образом, мы получаем общую плотность вероятности

$$\begin{aligned} \frac{1}{9} \left(\int_1^z \frac{dx_1}{9x_1} + \int_z^{10} \frac{10 dx_1}{9x_1} \right) &= \frac{1}{81} (\ln z - \ln 1 + 10 \ln 10 - 10 \ln z) = \\ &= \frac{10 \ln 10 - 9 \ln z}{81}. \end{aligned}$$

Вернемся теперь к поведению старших разрядов при сложении. Необходимо сделать предположения относительно величины сдвигов при выравнивании порядков перед сложением. Разумная гипотеза, которая, по-видимому, соответствует опыту, заключается в следующем:

- в половине случаев сдвига нет;
- в четверти случаев есть сдвиг на один разряд;
- в одной восьмой случаев сдвиг на два разряда;
- в одной шестнадцатой случаев сдвиг на три разряда и т. д.

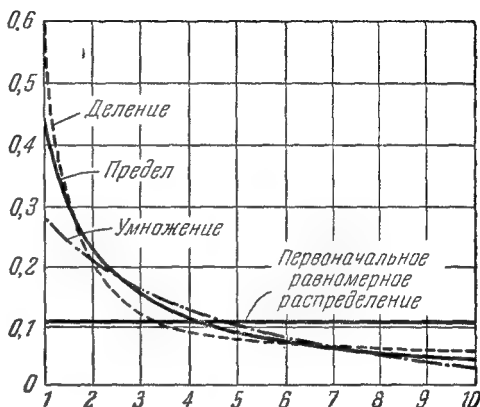


Рис. 2.10-1.

Если два числа взяты из равномерного распределения $1 \leq x < 10$, то в первом случае плотность суммы треугольная, идущая от нуля при $z=2$ к пику при $z=11$ и падающая опять до нуля при $z=20$ (рис. 2.10-2). Так как площадь должна быть 1, то мы имеем пик,

равный $\frac{1}{9}$. Все числа от 10 до 20 имеют старший разряд 1, и это выражается в площади

$$\frac{1}{2} + \frac{19}{81} = \frac{119}{162}.$$

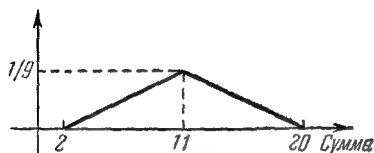


Рис. 2.10-2. Плотность распределения для суммы.

Очевидно, предположение равномерного распределения быстро приводит к преобладанию единиц в стар-

шем разряде. Не тратя больше времени и места на детальный анализ других случаев, укажем еще, что наличие сдвига изменит построенное распределение, но во всех случаях будет получаться преобладание старшей цифры 1. Таким образом, даже для сложения видно, что равномерное распределение старших разрядов — предположение нереалистичное.

Таким образом, в арифметике с плавающей запятой распространение ошибок через длинные цепи операций не так сурово, как это предсказывает модель равномерного распределения, так как благоприятное распределение старших разрядов влияет на распространение ошибки через операции умножения и деления.

Упражнения

2.10-1. Доказать второе из приведенных утверждений.

2.10-2. Рассмотреть теорию для операции вычитания.

§ 2.11. Анализ распространения ошибки при небольшом вычислении

Анализ распространения ошибки совсем прост в элементарных случаях, но в более громоздких он очень сложен. Элементарная теория основана на применении очевидного приближения

$$\delta f = \frac{df}{dx} \delta x,$$

и мы иногда будем его использовать.

Более сложные ситуации будут встречаться в различных местах этой книги, но общая трактовка выходит за пределы начального курса. Это высказывание не должно быть истолковано в том смысле, что сам предмет не важен.

§ 2.12. Общие замечания и библиография

Последние три теории, если их можно назвать теориями, пытаются исследовать, как растет округление в отдельных операциях. Такие теории, развитые дальше, чем это сделано здесь, необходимы, когда нужно сделать разумную оценку точности вычислений до того, как вычисление сделано, используя лишь общую структуру и размеры задачи. Поэтому в известном смысле различные фрагменты теорий, которые мы дали, не конкурируют, но скорее дополняют друг друга.

Вероятно, любая теория, в которой округление понимается в статистическом смысле и не учитывается распределение старших разрядов, не является достаточно точной. Неравномерное распределение старших разрядов обычно не обсуждается и не повлияло на развитие численного анализа, хотя известно давно. Таким образом, хорошей библиографии по этому вопросу нет.

Новичок в вычислительной работе всегда считает, что если ему везет, то он может пренебрегать эффектами округления; поэтому он старается игнорировать теории округления. И если он счастливый, то он действительно может это делать; но когда задачи становятся больше и запутаннее и когда от его результатов начинает многое зависеть, он постепенно вынужден думать о предмете, который в настоящее время развивается беспорядочно и убого. Будем надеяться, что лучшие и более полные теории будут найдены в недалеком будущем.

Несомненно, когда человек столкнется с необходимостью оценки точности задуманного вычисления до того, как затрачены деньги на программирование, отладку и т. д., он вынужден будет подумать о двух последних моделях округления и, вероятно, захочет, чтобы они были более развиты. Опыт работы показывает, что ошибки округления слишком часто коррелированы, допущение их независимости несправедливо, а значит, ограничиться продвижением лишь в направлениях, намеченных рассмотренными выше теориями, не удастся.

ГЛАВА 3

ИСЧИСЛЕНИЕ СУММ

§ 3.1. Введение и система обозначений

Исчисление разностей, введенное в гл. 1, было затем в гл. 2 использовано при рассмотрении важного вопроса об ошибках округления. В этой главе мы возвращаемся к главной теме первой части и беремся за исчисление сумм. Исчисление сумм связано с исчислением

разностей, как интегральное исчисление с дифференциальным. Здесь, как и в интегральном исчислении, нахождение обратного оператора, в сущности, основано на догадке.

Самым удобным обозначением для исчисления сумм является определение $\sum_{x=a}^b f(x)$ как суммы

$$f(a) + f(a+1) + \dots + f(b-1).$$

Это обозначение (заметим, что во всей этой главе мы считаем $h=1$) применяется Булем [3], Жорданом [19] и многими другими *); однако оно не является употребительным в других областях математики, и использование его могло бы привести к путанице. По-видимому, лучше все же иметь дело с затруднениями, которые возникают от применения неудобного, но общепринятого обозначения

$$\sum_{x=a}^b f(x) = f(a) + f(a+1) + \dots + f(b).$$

Методы суммирования будут целиком основываться на использовании прямого разностного оператора

$$\Delta f(x) = f(x+1) - f(x),$$

а не обратного разностного оператора ∇ и не центрального разностного оператора. Просуммировав последнее равенство от $x=a$ до $x=b-1$, получим

$$\sum_{x=a}^{b-1} \Delta f(x) = f(b) - f(a). \quad (3.1-1)$$

Это соответствует равенству

$$\int_a^b \frac{df(x)}{dx} dx = f(b) - f(a)$$

в интегральном исчислении.

Основная теорема исчисления сумм состоит в том, что если две функции, определенные на дискретном множестве точек, имеют одни и те же первые разности, то они различаются не более чем постоянным слагаемым. Это наводит на мысль о неопределенной сумме, соответствующей неопределенному интегралу, и аддитивной константе в таблице неопределенных сумм.

*) Тьюки предлагал обозначение $\sum_{x=a}^{<b}$ как более удобное.

В исчислении бесконечно малых таблица неопределенных интегралов основывается на соответствующей таблице производных; таким же образом таблица неопределенных сумм основывается на таблице разностей. Из

$$\Delta x^{(n)} = nx^{(n-1)},$$

применяя (3.1-1), получим

$$\sum_{x=a}^{b-1} x^{(n)} = \frac{b^{(n+1)} - a^{(n+1)}}{n+1} \quad (n \neq -1). \quad (3.1-2)$$

Для примера положим $n=0$, тогда

$$\sum_a^{b-1} x^{(0)} = \sum_a^{b-1} 1 = \frac{b^{(1)} - a^{(1)}}{1} = b - a.$$

Используя общую формулу (3.1-2) и очевидную линейность оператора \sum , мы можем находить суммы многочленов путем простого превращения степеней x в факториалы или при помощи чисел Стирлинга второго рода (§ 1.8), или повторяя деление многочленов (§ 1.6). Этим методом можно показать, что

$$\left. \begin{aligned} \sum_{x=1}^n x &= \sum_1^n x^{(1)} = \frac{(n+1)n}{2}, \\ \sum_{x=1}^n x^2 &= \sum_1^n [x^{(2)} + x^{(1)}] = \frac{(n+1)n(2n+1)}{6}, \\ \sum_{x=1}^n x^3 &= \sum_1^n [x^{(3)} + 3x^{(2)} + x^{(1)}] = \left[\frac{(n+1)n}{2} \right]^2, \\ \sum_{x=1}^n x^4 &= \frac{(n+1)n(2n+1)}{6} \cdot \frac{3n^2+3n-1}{5}, \\ \sum_{x=1}^n x^5 &= \left[\frac{(n+1)n}{2} \right]^2 \frac{2n^2+2n-1}{3}, \\ \sum_{x=1}^n x^6 &= \frac{(n+1)n(2n+1)}{6} \cdot \frac{3n^4+6n^3-3n+1}{7}, \\ \sum_{x=1}^n x^7 &= \left[\frac{(n+1)n}{2} \right]^2 \frac{3n^4+6n^3-n^2-4n+2}{6}. \end{aligned} \right\} \quad (3.1-3)$$

Упражнения

3.1-1. Проверить формулы для x^3 и x^5 , применяя метод деления многочленов.

3.1-2. Используя равенство $\Delta C(n, k) = C(n+1, k) - C(n, k) = C(n, k-1)$, показать, что

$$\sum_{x=k}^m C(x, k) = C(m+1, k+1).$$

3.1-3. Показать, что $\sum_{x=1}^n (2x-1) = n^2$, $\sum_{x=1}^n (2x-1)^2 = \frac{n(4n^2-1)}{3}$;

$$\sum_{x=1}^n (2x-1)^3 = n^2(2n^2-1).$$

§ 3.2. Формулы суммирования

Формула суммирования для $x^{(n)}$ верна также для отрицательных показателей ($n \neq -1$); например,

$$\sum_{x=1}^m \frac{1}{x(x+1)} = \sum_{x=1}^m (x-1)^{(-2)} = \frac{m^{(-1)} - 0^{(-1)}}{-1} = -\frac{1}{m+1} + \frac{1}{1} = \frac{m}{m+1}. \quad (3.2-1)$$

Подобным же образом,

$$\begin{aligned} \sum_{x=1}^m \frac{1}{x(x+1)(x+2)} &= \sum_{x=1}^m (x-1)^{(-3)} = \frac{m^{(-2)} - 0^{(-2)}}{-2} = \\ &= \frac{1}{2} \left[\frac{1}{1 \cdot 2} - \frac{1}{(m+1)(m+2)} \right]. \end{aligned}$$

Разностная формула

$$\Delta a^x = (a-1)a^x$$

приводит к суммированию геометрической прогрессии

$$\sum_0^m a^x = \frac{a^{m+1} - 1}{a - 1}.$$

Формулы для разностей синуса и косинуса

$$\Delta \sin(ax+b) = 2 \sin \frac{a}{2} \cos \left[a \left(x + \frac{1}{2} \right) + b \right],$$

$$\Delta \cos(ax+b) = -2 \sin \frac{a}{2} \sin \left[a \left(x + \frac{1}{2} \right) + b \right]$$

приводят к полезным формулам

$$\begin{aligned} \sum_{x=0}^m \sin(ax+b) &= -\frac{\cos\left[a\left(m+\frac{1}{2}\right)+b\right]-\cos\left[-\frac{a}{2}+b\right]}{2\sin\left(\frac{a}{2}\right)} = \\ &= \frac{\sin\left[\frac{a(m+1)}{2}\right]\sin\left[\frac{am}{2}+b\right]}{\sin\left(\frac{a}{2}\right)}, \quad (3.2-2), \end{aligned}$$

$$\begin{aligned} \sum_{x=0}^m \cos(ax+b) &= \frac{\sin\left[a\left(m+\frac{1}{2}\right)+b\right]-\sin\left[-\frac{a}{2}+b\right]}{2\sin\frac{a}{2}} = \\ &= \frac{\sin\left[\frac{a(m+1)}{2}\right]\cos\left[\frac{am}{2}+b\right]}{\sin\frac{a}{2}}. \quad (3.2-3). \end{aligned}$$

Можно ввести еще много других аналогичных формул, но мы не будем обременять ими читателя.

Упражнения

3.2-1. Вычислить $1+2+4+8+16+\dots+2^{24}$.

3.2-2. Показать, что

$$\sum_{x=0}^{2N-1} \cos \frac{\pi x}{N} = 0, \quad \sum_{x=0}^{2N-1} \sin \frac{\pi x}{N} = 0.$$

3.2-3. Показать, что

$$\begin{aligned} \sum_{x=0}^{2N-1} \cos \frac{\pi kx}{N} \cos \frac{\pi mx}{N} &= N\delta(m, k), \quad (0 < m+k < 2N); \\ \sum_{x=0}^{2N-1} \cos \frac{\pi kx}{N} \sin \frac{\pi mx}{N} &= 0; \\ \sum_{x=0}^{2N-1} \sin \frac{\pi kx}{N} \sin \frac{\pi mx}{N} &= N\delta(m, k), \quad (0 < m+k < 2N). \end{aligned}$$

3.2-4. Вычислить

$$\sum_{x=1}^N \frac{1}{x(x+1)(x+2)(x+3)}.$$

3.2-5. Вычислить

$$\sum_{m=0}^{N-1} \sin^m \theta \quad (\theta \neq 0, \pm\pi, \pm 2\pi, \dots).$$

$$\text{О т в е т: } \frac{(1 - \sin^N \theta)}{(1 - \sin \theta)}.$$

§ 3.3 Суммирование по частям

При интегрировании, кроме таблицы интегралов, применяются два метода:

- 1) замена переменного,
- 2) интегрирование по частям.

Первый из этих двух методов непригоден для исчисления сумм, зависящих от равноотстоящих аргументов. Это делает вычисление сумм в аналитическом конечном виде, вообще говоря, более трудным, чем вычисление интегралов.

С другой стороны, в исчислении сумм есть очень сильный метод, аналогичный интегрированию по частям. Интегрирование по частям основывается на формуле для производной от произведения

$$d(uv) = u dv + v du.$$

Из этого равенства формула для интегрирования по частям

$$\int u dv = uv - \int v du$$

находится интегрированием. Аналогично из формулы для разности произведения

$$\Delta(uv) = u \Delta v + v(x+1) \Delta u$$

суммированием получаем

$$\sum_{x=0}^{m-1} u \Delta v = u(m) v(m) - u(0) v(0) - \sum_{x=0}^{m-1} v(x+1) \Delta u(x).$$

Можно выбрать v так, чтобы $v(0) = 0$ (или любому другому заданному значению).

В качестве примера применения суммирования по частям рассмотрим выражение

$$\begin{aligned} \sum_{x=0}^{m-1} x a^x &= \left. \frac{x a^x}{a-1} \right|_0^m - \sum_{x=0}^{m-1} \frac{a^{x+1}}{a-1} \cdot 1 = \frac{x a^x}{a-1} - \left. \frac{a^{x+1}}{(a-1)^2} \right|_0^m = \\ &= \frac{a}{(1-a)^2} [(m-1) a^m - m a^{m-1} + 1]. \end{aligned}$$

Вообще применение суммирования по частям сильно напоминает применение интегрирования по частям. Например, суммирование

$$\sum \frac{1}{2^x} \sin \theta x$$

выполняется двукратным применением суммирования по частям и приведением подобных членов. В этих преобразованиях нет новых идей, а только скучная алгебра.

Упражнения

3.3-1. Показать, что

$$\sum_{x=1}^n \frac{1}{2^x} \sin \theta x = \frac{2 \sin \theta - 2^{-n} [2 \sin \theta (n+1) - \sin \theta n]}{1 - 8 \sin^2 \frac{\theta}{2}}.$$

3.3-2. Вычислить

$$\sum_{x=1}^n x^2 \cos x.$$

§ 3.4. Общие замечания

Лишь немногие конечные ряды можно просуммировать и представить компактной формулой. С другой стороны, неожиданно часто суммируются некоторые специальные ряды, например, содержащие биномиальные коэффициенты. Советуем читателю, прежде чем обращаться к вычислению ряда с помощью машины, попытаться просуммировать его руками. Удачное суммирование часто приводит к объяснению первоначальной задачи. Специальные методы суммирования рядов слишком многочисленны, чтобы рассматривать их здесь. Вероятно, лучшим справочником по суммированию рядов является Жолли [18]. См. также [1].

ГЛАВА 4

ВЫЧИСЛЕНИЕ БЕСКОНЕЧНЫХ РЯДОВ

§ 4.1. Введение

В большинстве книг о бесконечных рядах на многих страницах рассматриваются сходимость, расходимость и «суммируемость» рядов, но почти все они совершенно пренебрегают действительным вычислением (суммированием) рядов. Одна из причин этого состоит в том, что существует очень мало методов для суммирования рядов в конечном виде. Конечно, если возможно в конечном виде провести неопределенное суммирование и ряд сходится, то бесконечный ряд также

можно просуммировать, устремив верхний предел к бесконечности. В качестве примера мы имели (равенство (3.2-1))

$$\sum_{x=1}^n \frac{1}{x(x+1)} = 1 - \frac{1}{n+1}.$$

Отсюда

$$\sum_{x=1}^{\infty} \frac{1}{x(x+1)} = 1.$$

Вообще

$$\sum_{x=1}^{\infty} \frac{1}{x(x+1) \dots (x+k-1)} = \frac{1}{k-1} \cdot \frac{1}{(k-1)!} \quad (k \geq 2). \quad (4.1-1)$$

Задачу анализа нередко можно свести к вычислению бесконечного ряда. При этом часто требуется больше работы, чтобы вычислить ряд, чем чтобы решить первоначальную задачу; но иногда представление рядом является преимуществом. Дело в том, что, суммируя конечное число членов ряда, легче следить за точностью вычислений, чем действуя иным способом. Так, для вычисления

$$\int_0^x e^{-t^2} dt$$

для значений x , меньших 1, можно разложить экспоненту в бесконечный ряд. Интегрируя почленно, получим

$$\int_0^x e^{-t^2} dt = \int_0^x \sum_{k=0}^{\infty} \frac{(-1)^k t^{2k}}{k!} dt = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1) k!}.$$

Если мы интересуемся значениями x , меньшими 1, и хотим иметь восемь верных знаков, то достаточно взять 11 членов ряда. Действительно, ряд знакопеременный, его члены монотонны, он сходится и первый отброшенный член имеет знаменатель приблизительно $9,2 \cdot 10^8$. Если бы мы пытались вычислить значение интеграла каким-нибудь приближенным методом интегрирования, то задача оценки ошибки была бы более трудной.

Упражнения

4.1-1. Доказать формулу (4.1-1).

4.1-2. Написать ряд для

$$\text{Si}(x) = \int_0^x \frac{\sin t}{t} dt.$$

Сколько членов нужно взять для вычисления $\text{Si}(2\pi)$ с восемью десятичными знаками?

§ 4.2. Метод Куммера

Если данный ряд сходится быстро, то выбор способа вычисления его суммы не представляет затруднений. Если ряд сходится медленно, то мы ищем ряд с известной суммой, который сходится приблизительно с той же скоростью, что и данный ряд. Слова «приблизительно с той же скоростью» в действительности означают, что общий член разности этих двух рядов стремится к нулю быстрее, чем общий член исходного ряда. Найдя подходящий ряд, мы сводим задачу к вычислению суммы ряда, представляющего разность двух рядов; последний по определению сходится более быстро. В этом и состоит идея метода Куммера.

Пусть данный ряд есть

$$S = \sum_{x=r}^{\infty} a_x$$

и предположим, что мы знаем сумму

$$S' = \sum_{x=r}^{\infty} c_x a_x$$

где $c_x \rightarrow c$ при $x \rightarrow \infty$. Тогда

$$S = \frac{S'}{c} + \sum_{x=r}^{\infty} \left(1 - \frac{c_x}{c}\right) a_x.$$

В качестве примера рассмотрим ряд

$$S = \sum_{x=1}^{\infty} \frac{x}{(x^2+1)^2},$$

который сходится как $\frac{1}{x^3}$. Взяв

$$S' = \sum_{x=2}^{\infty} \frac{1}{(x-1)x(x+1)} \quad (c=1)$$

в качестве ряда для сравнения, при условии, что член с $x=1$ взят отдельно, получим

$$\begin{aligned} S &= S' + (S - S') = S' + \frac{1}{4} + \sum_2^{\infty} \left[\frac{x}{(x^2+1)^2} - \frac{1}{(x-1)x(x+1)} \right] = \\ &= \frac{1}{4} + \frac{1}{4} + \sum_2^{\infty} \frac{x^4 - x^2 - (x^4 + 2x^2 + 1)}{x(x^2-1)(x^2+1)^2} = \\ &= \frac{1}{2} - \sum_2^{\infty} \frac{3x^2 + 1}{x(x^2-1)(x^2+1)^2} = 0,39711677... \end{aligned}$$

Новый ряд сходится как $1/x^4$.

Упражнения

4.2-1. Использовать $\sum_1^{\infty} \frac{1}{x(x+1)}$ для приближения $\sum_1^{\infty} \frac{1}{x^2}$.

4.2-2. Приблизить результат упражнения 4.2-1 с помощью

$$\sum_1^{\infty} \frac{1}{x(x+1)(x+2)^*},$$

Продолжать до k -го шага.

§ 4.3. Некоторые специальные суммы

Метод Куммера требует рядов для сравнения и знания их сумм. Одна из самых полезных последовательностей рядов для сравнения, кроме рядов (4.1-1), есть последовательность сумм

$$S_k = \sum_{x=1}^{\infty} \frac{1}{x^k} \quad (k=2, 3, \dots), \quad (4.3-1)$$

являющихся значениями дзета-функции Римана

$$\zeta(z) = \sum_{n=1}^{\infty} \frac{1}{n^z}$$

(см. таблицу 8.7-1 для $z=2, 3, \dots, 17$). Значения для четных целых z известны в конечном виде,

$$\begin{aligned} S_2 = \zeta(2) &= \frac{\pi^2}{6}; & S_4 &= \frac{\pi^4}{90}; \\ S_6 &= \frac{\pi^6}{945}; & S_8 &= \frac{\pi^8}{9450}, \end{aligned}$$

но для нечетных чисел конечный вид неизвестен.

В многочисленных книгах и таблицах *) имеется много других рядов, имеющих известные суммы, и читатель, несомненно, знаком с теми, которые возникают из разложения элементарных функций в ряд Маклорена.

§ 4.4. Метод Эйлера

Другим методом численного суммирования рядов является метод Эйлера. К нему можно прийти следующим образом.

Рассмотрим конечный ряд

$$\sum_{k=0}^{n-1} a_k t^k. \quad (4.4-1)$$

*) См. [18] и [1].

Применим формулу суммирования по частям

$$\sum_a^{b-1} u(k) \Delta v(k) = [u(k)v(k)]_a^b - \sum_a^{b-1} v(k+1) \Delta u(k).$$

Положим $u(k) = a_k$, $\Delta v(k) = t^k$. Так как мы можем использовать любую аддитивную постоянную, то возьмем

$$v(k) = \sum_{x=0}^{k-1} t^x = \frac{1-t^k}{1-t}.$$

Тогда

$$\begin{aligned} \sum_{k=0}^{n-1} a_k t^k &= \left[a_k \frac{1-t^k}{1-t} \right]_0^n - \sum_{k=0}^{n-1} \frac{1-t^{k+1}}{1-t} \Delta a_k = \\ &= a_n \left(\frac{1-t^n}{1-t} \right) - \frac{1}{1-t} \sum_0^{n-1} \Delta a_k + \frac{t}{1-t} \sum_0^{n-1} t^k \Delta a_k; \quad (4.4-2) \end{aligned}$$

но

$$\sum_{k=0}^{n-1} \Delta a_k = (a_1 - a_0) + (a_2 - a_1) + \dots + (a_n - a_{n-1}) = a_n - a_0,$$

так что (4.4-2) принимает вид

$$\begin{aligned} \sum_{k=0}^{n-1} a_k t^k &= a_n \left(\frac{1-t^n}{1-t} \right) - \frac{a_n - a_0}{1-t} + \frac{t}{1-t} \sum_0^{n-1} t^k \Delta a_k = \\ &= \frac{a_0}{1-t} - \frac{a_n t^n}{1-t} + \frac{t}{1-t} \sum_{k=0}^{n-1} t^k \Delta a_k. \quad (4.4-3) \end{aligned}$$

Мы применяем суммирование по частям к третьему члену, заметив, что он имеет тот же вид, что и исходный ряд, только a_k заменено на Δa_k в (4.4-3):

$$\frac{t}{1-t} \sum_{k=0}^{n-1} t^k \Delta a_k = \frac{t}{1-t} \left(\frac{\Delta a_0}{1-t} - \frac{\Delta a_n t^n}{1-t} + \frac{t}{1-t} \sum_{k=0}^{n-1} t^k \Delta^2 a_k \right).$$

Таким образом, в результате применения суммирования по частям дважды получим

$$\sum_{k=0}^{n-1} a_k t^k = \frac{a_0}{1-t} + \frac{t \Delta a_0}{(1-t)^2} + \frac{t^2}{(1-t)^2} \sum_{k=0}^{n-1} t^k \Delta^2 a_k - \frac{a_n t^n}{1-t} - \frac{1}{1-t} \frac{\Delta a_n t^{n+1}}{1-t}.$$

После r таких преобразований выражение (4.4-1) примет вид

$$\sum_{k=0}^{n-1} a_k t^k = \frac{1}{1-t} \sum_{i=0}^{r-1} \left(\frac{t}{1-t}\right)^i \Delta^i a_0 + \left(\frac{t}{1-t}\right)^r \sum_{k=0}^{n-1} t^k \Delta^r a_k - \\ - \frac{t^n}{1-t} \sum_{i=0}^{r-1} \left(\frac{t}{1-t}\right)^i \Delta^i a_n.$$

Так как первоначальный ряд сходится, то для данного $\varepsilon > 0$ существует такое n_0 , что при $n > n_0$ справедливо $|a_n| < \frac{\varepsilon}{2^r}$. Следовательно, по (1.4-3), $|\Delta^r a_n| \leq \varepsilon$. Последний член в полученном выражении стремится поэтому к нулю при $n \rightarrow \infty$, и мы имеем

$$\sum_{k=0}^{\infty} a_k t^k = \frac{1}{1-t} \sum_{i=0}^{r-1} \left(\frac{t}{1-t}\right)^i \Delta^i a_0 + \left(\frac{t}{1-t}\right)^r \sum_{k=0}^{\infty} t^k \Delta^r a_k. \quad (4.4-4)$$

Предположим, что члены ряда (4.4-1) изменяются достаточно гладко, так что второе слагаемое справа стремится к нулю при $r \rightarrow \infty$. Тогда остается ряд

$$\sum_{k=0}^{\infty} a_k t^k = \frac{1}{1-t} \sum_{i=0}^{\infty} \left(\frac{t}{1-t}\right)^i \Delta^i a_0. \quad (4.4-5)$$

Рассмотрим пример применения указанного метода. Если дан ряд $\sum_{k=0}^{\infty} u_k$, у которого $\frac{u_{k+1}}{u_k}$ стремится к t , то можно написать

$$\sum_0^{\infty} u_k = \sum_0^{\infty} a_k t^k,$$

где $\frac{a_{k+1}}{a_k}$ стремится к 1, и применить метод Эйлера.

Наиболее часто этот метод применяется для $t = -1$. По (4.4-5) находим

$$\sum_{k=0}^{\infty} (-1)^k a_k = \frac{1}{2} \sum_{i=0}^{\infty} \frac{(-1)^i}{2^i} \Delta^i a_0. \quad (4.4-6)$$

Иногда преобразование Эйлера делает ряд сходящимся быстрее, но иногда и нет. Рассмотрим следующие примеры.

Пример 1.

$$\sum_{k=0}^{\infty} \frac{(-1)^k}{2^k}.$$

Мы имеем

$$a_n = \frac{1}{2^n}, \quad \Delta a_n = \frac{1}{2^{n+1}} - \frac{1}{2^n} = \frac{-1}{2^{n+1}}, \quad \Delta^i a_0 = \frac{(-1)^i}{2^i}$$

и, таким образом, по (4.4-6) получаем ряд

$$\sum_{k=0}^{\infty} \frac{(-1)^k}{2^k} = \frac{1}{2} \sum_{i=0}^{\infty} \frac{(-1)^i}{2^i} \cdot \frac{(-1)^i}{2^i} = \frac{1}{2} \sum_{i=0}^{\infty} \frac{1}{4^i},$$

который сходится быстрее первоначального.

Пример 2.

$$\sum_{k=0}^{\infty} \frac{(-1)^k}{3^k}.$$

Легко видеть, что

$$\Delta^i a_0 = \frac{(-2)^i}{3^i}.$$

Таким образом, преобразование Эйлера дает ряд

$$\sum_{k=0}^{\infty} \frac{(-1)^k}{3^k} = \frac{1}{2} \sum_{i=0}^{\infty} \frac{(-1)^i}{2^i} \cdot \frac{(-2)^i}{3^i} = \frac{1}{2} \sum_{i=0}^{\infty} \frac{1}{3^i},$$

который сходится несколько медленнее, чем заданный.

Пример 3. Аналогично

$$\sum_{k=0}^{\infty} \frac{(-1)^k}{4^k} = \frac{1}{2} \sum_{i=0}^{\infty} \left(\frac{3}{8}\right)^i;$$

последний ряд сходится более медленно.

«Точка перелома» $|a_i|$ для применения метода Эйлера к такому знакопеременному ряду, по-видимому, лежит между $1/2$ и $1/3$.

На практике обычно суммируют первые несколько (скажем, 10) членов непосредственно и применяют преобразование Эйлера к оставшимся *)

Упражнения

4.4-1. Применить метод Эйлера к ряду

$$\ln 2 = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} = \sum_{k=1}^{\infty} \frac{1}{k2^k}.$$

*) См. [4], а также J. B. Rosser, Transformations to Speed the Convergence of Series, J. Research. Nat. Bur. Standards, vol. 46, 1951.

4.4.2. Показать, что

$$\operatorname{arctg} 1 = \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} = \frac{1}{2} \sum_{k=0}^{\infty} \frac{k!}{1 \cdot 3 \cdot 5 \dots (2k+1)}.$$

4.4.3. Сложить первые восемь членов ряда для $\ln 2$ и применить метод Эйлера к оставшимся.

Ответ: $1 - 1/2 + 1/3 - \dots - 1/8 = 0,63452381$

оставшаяся часть равна $\frac{1}{2} \cdot 0,11724074$.

§ 4.5. Нелинейное преобразование

Большинство преобразований, применяемых в математике, линейны, но нелинейное преобразование частичных сумм ряда

$$T(S_n) = \frac{S_{n+1}S_{n-1} - S_n^2}{S_{n+1} - 2S_n + S_{n-1}} \quad (4.5-1)$$

часто очень полезно. Если бы T было линейным, мы имели бы

$$T(CS_n) = CT(S_n)$$

и

$$T(S_n + U_n) = T(S_n) + T(U_n).$$

Второе условие, вообще говоря, неверно. Однако справедливо более слабое равенство

$$T(S_n + C) = T(S_n) + C \quad (C - \text{const}).$$

Это преобразование полезно в тех случаях, когда ряд ведет себя, грубо говоря, как геометрическая прогрессия. В качестве иллюстрации эффективности преобразования рассмотрим ряд Лейбница

$$\pi = 4 - 4/3 + 4/5 - 4/7 + \dots$$

Сходимость его так медленна, что практически незаметна, но одно применение преобразования дает следующее:

n	S_n	$T(S_n)$	n	S_n	$T(S_n)$
0	4,00000		5	2,97605	3,14271
1	2,66667	3,16667	6	3,28374	3,14088
2	3,46667	3,13333	7	3,01707	3,14207
3	2,89524	3,14524	8	3,25237	3,14125
4	3,33968	3,13968	9	3,04184	

Теория этого преобразования и многих родственных ему, которые часто даже более сильны, хорошо изучена и не будет рассматриваться здесь.

Это преобразование полезно не только для обработки медленно сходящихся и расходящихся рядов, но также и для многих итерационных последовательностей, так же как метод Ньютона для нахождения нулей функций. Если сходимость имеет в каком-то смысле скорость геометрической прогрессии, то преобразование часто улучшает дело.

Упражнение 4.5-1. Применяя преобразование (4.5-1), вычислить

$$\sum_{k=1}^{\infty} \left[\frac{(-1)^{k-1}}{k} \right] \text{ до четвертого десятичного знака.}$$

§ 4.6. Степенные ряды

Степенные ряды широко применяются в математике, особенно в линейных задачах. Но даже в нелинейных задачах степенные ряды — полезный инструмент. Так как в нелинейных задачах вычисление последовательных коэффициентов часто очень трудоемко, вычислительная машина должна находить их сама. Для данного числа коэффициентов легко написать программы сложения, вычитания, умножения, деления, подстановки одного степенного ряда в другой и обращения ряда.

Преимущество использования степенных рядов или многочленов, как в задаче § 1.9, в том, что большую часть действий над коэффициентами можно выполнить один раз перед началом вычислений.

Абстрактное математическое описание того, что было сделано в задаче § 1.9, выглядит так: 11 точек могут рассматриваться как одна точка в 11-мерном пространстве. Первая операция получения факториального многочлена Ньютона преобразовала точку исходного пространства в соответствующую точку 11-мерного пространства коэффициентов. Использование чисел Стирлинга для вычисления непосредственного многочленного представления было равносильно изменению базиса пространства коэффициентов. Последующие операции были преобразованиями в пространстве коэффициентов, тогда как окончательное вычисление было преобразованием обратно в исходное пространство. Таким образом, большинство действий производилось в преобразованном пространстве, а не в исходном пространстве данных и ответа. Это характерно для метода степенных рядов; мы оперируем с коэффициентами разложения и возвращаемся к данному пространству только в конце. Подобные замечания относятся как к асимптотическим рядам, так и к методам, обсуждающимся в этой главе.

Упражнения

4.6-1. Составить блок-схему для программы умножения двух степенных рядов с k членами, чтобы получить k членов в результате.

4.6-2. То же самое для деления.

4.6-3. То же самое для подстановки одного ряда в другой.

4.6-4. К данному интегралу

$$g(T) = \int_0^{\infty} \frac{f(x) \left(\frac{x}{T}\right)^2 e^{x/T}}{\left(e^{x/T} - 1\right)^2} dx$$

применить метод § 1.9 и принять $f(x) = x^k$. Показать, что $g(T) = (n+2)! T^{n+1} \zeta(n+2)$. Вывести отсюда, что если $g(T)$ может быть аппроксимирована

$$g(T) = \sum_{n=0}^N a_n T^{n+1},$$

то

$$f(x) = \sum_{n=0}^N \frac{a_n x^n}{(n+2)! \zeta(n+2)}.$$

§ 4.7. Разложение по специальным функциям

Кроме разложения в степенные ряды, в анализе часто применяется разложение по специальным функциям, таким как полиномы Лежандра $P_n(x)$, полиномы Лагерра $L_n(x)$, полиномы Эрмита $H_n(x)$, функции Бесселя $J_n(x)$ и т. д. При поверхностном рассмотрении может показаться, что эти разложения бесполезны из-за трудности вычисления значений самих специальных функций. Однако известно, что большинство семейств специальных функций удовлетворяет трехчленному рекуррентному соотношению вида

$$f_{n+1}(x) = A(n, x) f_n(x) + B(n) f_{n-1}(x).$$

В тех случаях, когда специальные функции суть многочлены, многочлены нулевого и первого порядков особенно легко вычислять для каждого значения x . Таким образом, работа с разложением в ряд по специальным функциям не больше, чем работа по вычислению степенного ряда.

Следует обратить внимание на накопление ошибки при использовании рекуррентного соотношения для вычисления последовательных функций; но обычно ошибка не слишком быстро растет для умеренных (скажем, порядка 15) значений индекса n .

§ 4.8. Интегралы как приближения сумм

Определенный интеграл определяется как предел суммы. Поэтому интегралы могут быть использованы для приближенного вычисления сумм. Одна из формул такого типа указывается в § 12.3 и может

быть записана так:

$$f_0 + f_1 + \dots + f_n = \int_0^n f(x) dx + \frac{1}{2} f_0 + \frac{1}{2} f_n - \\ - \frac{1}{12} (\Delta f_0 - \Delta f_{n-1}) + \frac{1}{24} (\Delta^2 f_0 + \Delta^2 f_{n-2}) - \\ - \frac{19}{720} (\Delta^3 f_0 - \Delta^3 f_{n-3}) + \frac{3}{160} (\Delta^4 f_0 + \Delta^4 f_{n-4}) + \dots \quad (4.8-1)$$

Родственной формулой является формула Эйлера — Маклорена ([43], стр 127—128)

$$f_0 + f_1 + \dots + f_n = \int_0^n f(x) dx + \frac{1}{2} f_0 + \frac{1}{2} f_n - \frac{1}{12} [f'_0 - f'_n] + \\ + \frac{1}{720} [f'''_0 - f'''_n] - \frac{1}{30240} [f^{(5)}_0 - f^{(5)}_n] + \dots \quad (4.8-2)$$

§ 4.9. Дигамма-функция

В заключение этой главы вспомним формулу

$$\int x^m dx = \frac{x^{m+1}}{m+1} + C,$$

которая неприменима, когда $m = -1$. Действительно, при $m = -1$ этот интеграл определяет новую функцию $\ln x$. Аналогично формула суммирования

$$\sum_{t=1}^{x-1} t^{(m)} = \frac{x^{(m+1)}}{m+1} + C$$

неприменима при $m = -1$ и соответственно определяет новую функцию, названную *дигамма-функцией* и обозначенную $F(x)$

$$F(x) = \sum_{r=1}^x \frac{1}{r} - \gamma, \quad (4.9-1)$$

где $\gamma = 0,5772156649 \dots$ (константа Эйлера) и, следовательно, $F(0) = -\gamma$.

Эта формула применяется, когда x — целое число. Другой вид

$$F(x) = \sum_{r=1}^{\infty} \frac{x}{r(r+x)} - \gamma, \quad F(0) = -\gamma \quad (4.9-2)$$

дает возможность распространить определение $F(x)$ на нецелые значения. Надо показать, следовательно, что это новое определение для

целых x совпадает со старым. Для этого достаточно установить, что $F(x)$ удовлетворяет уравнению

$$\Delta F(x) = \frac{1}{x+1}.$$

(Аналогично при доказательстве того, что функция удовлетворяет формуле интегрирования, проверяют, что ее производная есть подынтегральная функция.) Применяя (4.9-2), находим

$$\begin{aligned} \Delta F(x) &= \sum_{r=1}^{\infty} \left[\frac{x+1}{r(r+x+1)} - \frac{x}{r(r+x)} \right] = \\ &= \sum_{r=1}^{\infty} \left[\frac{1}{(r+x)(r+x+1)} \right] = \sum_{r=1}^{\infty} \left[\frac{1}{r+x} - \frac{1}{r+x+1} \right] = \frac{1}{x+1}. \end{aligned}$$

Между дигамма-функцией и натуральным логарифмом существует следующее соотношение:

$$F(x) = \lim_{n \rightarrow \infty} \left(\ln(x+n+1) - \frac{1}{x+1} - \frac{1}{x+2} - \dots - \frac{1}{x+n} \right).$$

Есть другая связь между известными функциями

$$\frac{d}{dx} \ln [\Gamma(1+x)] = F(x).$$

Дальнейшим дифференцированием мы получим так называемую тригамма-функцию

$$\frac{d^2}{dx^2} \ln [\Gamma(1+x)] = \sum_1^{\infty} \frac{1}{(r+x)^2} = F'(x),$$

тетрагамма-функцию

$$\frac{d^3}{dx^3} \ln [\Gamma(1+x)] = -2 \sum_1^{\infty} \frac{1}{(r+x)^3} = -F''(x),$$

пентагамма-функцию

$$\frac{d^4}{dx^4} \ln [\Gamma(1+x)] = 6 \sum_1^{\infty} \frac{1}{(r+x)^4} = F'''(x),$$

и т. д.

Эти функции протабулированы [35] и могут применяться для суммирования рядов, общий член которых есть рациональная функция. Для примера предположим, что мы имеем ряд

$$S = \sum_{x=1}^{\infty} \frac{P_{n-2}(x)}{P_n(x)},$$

где $P_n(x)$ — многочлен степени n и $P_{n-2}(x)$ по крайней мере на две степени меньше, чем $P_n(x)$. Используя элементарные дроби, мы можем написать (считая, что a_i есть нули $P_n(x)$)

$$S = \sum_{x=1}^{\infty} \left[\frac{A_1}{x-a_1} + \frac{A_2}{x-a_2} + \dots + \frac{A_k}{x-a_k} + \frac{B_1}{(x-a_1)^2} + \frac{B_2}{(x-a_2)^2} + \dots \right. \\ \left. \dots + \frac{B_k}{(x-a_k)^2} + \dots + \frac{M_1}{(x-a_1)^m} + \frac{M_2}{(x-a_2)^m} + \dots + \frac{M_k}{(x-a_k)^m} \right].$$

Здесь нельзя отделить первую группу членов, так как отдельные ряды расходятся. Однако легко видеть, что $\sum_{i=1}^k A_i = 0$, следовательно, можно написать

$$\left[\left(\frac{A_1}{x-a_1} - \frac{A_1}{x} \right) + \left(\frac{A_2}{x-a_2} - \frac{A_2}{x} \right) + \dots + \left(\frac{A_k}{x-a_k} - \frac{A_k}{x} \right) \right],$$

не меняя значения суммы группы. Теперь мы имеем ряд в виде, в котором он может быть легко просуммирован. В качестве примера рассмотрим (см. [35])

$$S = \sum_{x=1}^{\infty} \frac{1}{(4x+2)(4x+1)(4x+3)^2} = \\ = \sum_{x=1}^{\infty} \left[-\frac{1}{4x+2} + \frac{1}{4} \cdot \frac{1}{4x+1} + \frac{3}{4} \cdot \frac{1}{4x+3} + \frac{1}{2} \cdot \frac{1}{(4x+3)^2} \right] = \\ = \frac{1}{4} F\left(\frac{1}{2}\right) - \frac{1}{16} F\left(\frac{1}{4}\right) - \frac{3}{16} F\left(\frac{3}{4}\right) + \frac{1}{32} F\left(\frac{3}{4}\right).$$

ГЛАВА 5

УРАВНЕНИЯ В КОНЕЧНЫХ РАЗНОСТЯХ

§ 5.1. Система обозначений

Разностные уравнения в исчислении разностей соответствуют дифференциальным уравнениям в дифференциальном исчислении. Так, например, для неизвестной функции y можно иметь дифференциальное уравнение

$$y' + 2y = x$$

или разностное уравнение

$$\Delta y + 2y = x. \quad (5.1-1)$$

Любопытно, что в этом виде аналогия между дифференциальными и разностными уравнениями не полная. Но если мы подставим

$$\Delta y = y(x+1) - y(x)$$

в (5.1-1), то получим уравнение

$$y(x+1) + y(x) = x. \quad (5.1-2)$$

Ясно, что всегда можно перейти от одного вида разностного уравнения к другому. Мы будем всегда пользоваться второй формой записи. В ней уравнение можно решить в основном так же, как мы решаем уравнение

$$y' + y = x.$$

Дело в том, что в этой форме более ясно раскрывается природа задачи. Так, если дано уравнение

$$\Delta^2 y + 2\Delta y + y = f(x),$$

то можно считать, что мы имеем разностное уравнение второго порядка. Но оно приводит к уравнению

$$f(x) = y(x+2) - 2y(x+1) + y(x) + 2[y(x+1) - y(x)] + y(x) = y(x+2),$$

которое тривиально. Максимальный порядок разности аргументов неизвестной функции определяет порядок разностного уравнения.

Так как между разностными и дифференциальными уравнениями есть прямая аналогия, стоит повторить некоторые детали того, как решаются дифференциальные уравнения. Как и раньше, все приемы, кроме замены независимого переменного, применяются в обоих случаях.

§ 5.2. Пример разностного уравнения первого порядка

Чтобы решить дифференциальное уравнение

$$y' + y = x, \quad (5.2-1)$$

сначала рассматривается однородное уравнение

$$y' + y = 0 \quad (5.2-2)$$

и ищется общее решение однородного уравнения. Мы предполагаем, что при должном выборе m решением однородного уравнения будет $y = e^{mx}$. Подставляя в уравнение, получаем

$$\begin{aligned} me^{mx} + e^{mx} &= 0, \\ e^{mx}(m+1) &= 0, \\ m+1 &= 0 \quad (e^{mx} \neq 0), \\ m &= -1, \end{aligned}$$

откуда

$$y = Ce^{-x}, \quad (5.2-3)$$

где C — произвольная постоянная.

Затем ищется частное решение полного уравнения. Принимая во внимание правую часть, положим,

$$y = ax + b.$$

Дифференциальное уравнение (5.2-1) дает

$$a + ax + b = x,$$

так что

$$a = 1, \quad b = -1$$

и

$$y = x - 1 \quad (5.2-4)$$

есть частное решение (5.2-1).

Итак, общее решение есть сумма (5.2-3) и (5.2-4):

$$y = Ce^{-x} + (x - 1),$$

т. е. сумма общего решения однородного уравнения и частного решения. Конечно, этот метод применим только к линейным уравнениям.

Для разностного уравнения

$$y(x+1) + y(x) = x$$

делается то же самое. В однородное уравнение

$$y(x+1) + y(x) = 0$$

мы подставляем вместо e^{mx}

$$y = p^x,$$

что является простым изменением обозначения, так как можно положить $e^m = p$. Получаем

$$p^x(p+1) = 0,$$

откуда

$$p = -1 \quad \text{и} \quad y = C(-1)^x.$$

Для полного уравнения мы опять предполагаем, что

$$y = ax + b.$$

Разностное уравнение дает

$$a(x+1) + b + ax + b = x, \quad 2a = 1, \quad a = 1/2,$$

так что $a + 2b = 0$, $b = -1/4$ и $y = x/2 - 1/4$. Тогда общее решение есть

$$y = C(-1)^x + \frac{x}{2} - \frac{1}{4}.$$

Таким образом, мы видим тесную связь линейных уравнений первого порядка различных типов.

§ 5.3. Пример уравнения второго порядка

Для линейных уравнений более высокого порядка применяется та же техника. Рассмотрим, например, числа Фибоначчи, которые определены уравнением

$$y_{n+1} = y_n + y_{n-1}$$

с

$$y_0 = 0, \quad y_1 = 1.$$

Прежде всего ищем общее решение. Пробуя

$$y = p^n,$$

получаем

$$p^2 - p - 1 = 0$$

или

$$p = \frac{1 \pm \sqrt{5}}{2}.$$

Таким образом,

$$y_n = C_1 \left(\frac{1 + \sqrt{5}}{2} \right)^n + C_2 \left(\frac{1 - \sqrt{5}}{2} \right)^n.$$

Условие $y_0 = 0$ дает

$$0 = C_1 + C_2,$$

а условие $y_1 = 1$:

$$y_1 = 1 = C_1 \left(\frac{1 + \sqrt{5}}{2} \right) + C_2 \left(\frac{1 - \sqrt{5}}{2} \right).$$

Из этих уравнений получаем

$$C_1 = \frac{1}{\sqrt{5}} = -C_2 \quad \text{и} \quad y_n = \frac{1}{\sqrt{5}} \left[\left(\frac{1 + \sqrt{5}}{2} \right)^n - \left(\frac{1 - \sqrt{5}}{2} \right)^n \right].$$

Этот пример выбран прежде всего потому, что само разностное уравнение дает возможность легко вычислять по очереди последовательные числа. Так

$$\begin{array}{llll} y_0 = 0, & y_3 = 2, & y_6 = 8, & y_9 = 34, \\ y_1 = 1, & y_4 = 3, & y_7 = 13, & y_{10} = 55, \\ y_2 = 1, & y_5 = 5, & y_8 = 21, & y_{11} = 89. \end{array}$$

Никакого сомнения в существовании решения не возникает, но общее решение, очевидно, слишком трудно применить на практике, хотя оно, может быть, полезно для вычисления некоторых отдельных чисел.

§ 5.4. Линейные разностные уравнения с постоянными коэффициентами

Продолжим аналогию с линейными дифференциальными уравнениями. В случае двойного корня характеристического уравнения мы ищем решения как в виде e^{mx} , так и xe^{mx} . Рассмотрим, например, уравнение

$$y'' - 2y' + y = 0.$$

Полагая $y = e^{mx}$, получаем

$$m^2 - 2m + 1 = 0$$

или

$$m_1 = 1, \quad m_2 = 1.$$

Тогда e^x и xe^x удовлетворяют уравнению и общее решение есть

$$y = C_1 e^x + C_2 x e^x.$$

Рассмотрим теперь разностное уравнение

$$y(x+2) - 2y(x+1) + y(x) = 0.$$

Подставляя $y = p^x$, получаем

$$p^2 - 2p + 1 = 0, \quad p_1 = 1, \quad p_2 = 1.$$

Тогда каждое из решений $(1)^x = 1$, $x(1)^x = x$ удовлетворяет уравнению и общее решение есть

$$y = C_1 + C_2 x.$$

Вспомним, что аналогичным образом мы действуем, когда правая часть неоднородного уравнения содержит член, который есть в решении однородного уравнения. В случае

$$y' + y = e^{-x}$$

однородное уравнение

$$y' + y = 0$$

имеет решение $y = Ce^{-x}$. Поэтому частное решение всего уравнения мы пытаемся найти в виде

$$y = kxe^{-x}, \quad ke^{-x} - kxe^{-x} + kxe^{-x} = e^{-x}, \quad k = 1.$$

Окончательно получаем

$$y = Ce^{-x} + xe^{-x} = (C + x)e^{-x}.$$

Точно так же, рассматривая разностное уравнение

$$y(x+1) - y(x) = 1,$$

мы составляем однородное уравнение

$$y(x+1) - y(x) = 0.$$

Оно имеет решение $y = C$, где C — произвольная постоянная. Поэтому, когда мы угадываем частное решение, то пробуем

$$y = kx.$$

Подставляя его, видим, что $k = 1$, следовательно, общее решение первоначального уравнения есть

$$y = C + x.$$

Милн-Томсон [27] рассматривает решение разностных уравнений более подробно. Он показывает, в частности, как знание одного решения однородного линейного разностного уравнения дает возможность понизить порядок уравнения на единицу.

§ 5.5. Пример

Цель этого параграфа — дать пример применения разностных уравнений и показать, как некоторые идеи и методы из области обыкновенных дифференциальных уравнений можно естественным образом применять к разностным уравнениям.

Задача ставится так: найти интегралы

$$I(n) = \int_0^{\infty} x^n e^{-x} \sin x \, dx \quad (n \geq 0),$$

$$K(n) = \int_0^{\infty} x^n e^{-x} \cos x \, dx$$

и, в частности, показать, что $I(4k+3) = 0$.

Интегрирование по частям дает систему разностных уравнений

$$\left. \begin{aligned} I(n) &= \frac{n}{2} [I(n-1) + K(n-1)], \\ K(n) &= \frac{n}{2} [-I(n-1) + K(n-1)]. \end{aligned} \right\} \quad (5.5-1)$$

Из стандартных таблиц интегралов находим $I(0) = K(0) = \frac{1}{2}$.

Эти разностные уравнения имеют переменные коэффициенты $\frac{n}{2}$; поэтому $I(n)$ имеет множитель n , $I(n+1)$ имеет множитель $(n+1)$ и т. д., так что $I(n)$ ведет себя, как $n!$. Таким образом, мы приходим к подстановке

$$I(n) = n! j(n), \quad K(n) = n! k(n).$$

При этом мы не учли множитель $\frac{1}{2}$ в коэффициентах разностных уравнений. Приняв во внимание и его, получаем преобразование

$$I(n) = \frac{n!}{2^n} j(n), \quad K(n) = \frac{n!}{2^n} k(n), \quad (5.5-2)$$

которое приводит к системе

$$\left. \begin{aligned} j(n) &= j(n-1) + k(n-1), & j(0) &= \frac{1}{2}, \\ k(n) &= -j(n-1) + k(n-1), & k(0) &= \frac{1}{2}. \end{aligned} \right\} \quad (5.5-3)$$

Если бы это была система обыкновенных дифференциальных уравнений, мы или пытались бы вспомнить, как решаются системы, или свели задачу к одному уравнению второго порядка, продифференцировав одно уравнение и исключив одну переменную. Последнее, по-видимому, легче приспособить к разностным уравнениям.

Напишем верхнее уравнение (5.5-3) в виде

$$j(n+1) = j(n) + k(n) \quad (5.5-4)$$

и, используя его во втором уравнении (5.5-3), исключим $k(n)$. Затем, используя первое из уравнений (5.5-3) для исключения $k(n-1)$, получаем

$$j(n+1) - 2j(n) + 2j(n-1) = 0. \quad (5.5-5)$$

Это — линейное разностное уравнение второго порядка с постоянными коэффициентами. Начальные условия находим из (5.5-3) и (5.5-4):

$$j(0) = \frac{1}{2}, \quad j(1) = 1. \quad (5.5-6)$$

Общее решение (5.5-5) есть (здесь $i = \sqrt{-1}$)

$$j(n) = C_1(1+i)^n + C_2(1-i)^n,$$

а начальные условия (5.5-6) дают

$$C_1 + C_2 = \frac{1}{2}, \quad C_1(1+i) + C_2(1-i) = 1,$$

или

$$C_1 = \frac{1-i}{4}, \quad C_2 = \frac{1+i}{4}.$$

Таким образом,

$$j(n) = \frac{1}{2} [(1+i)^{n-1} + (1-i)^{n-1}]$$

и

$$I(n) = \frac{n!}{2^{n+1}} [(1+i)^{n-1} + (1-i)^{n-1}]$$

(на самом деле $I(n)$ и $j(n)$, как видно из полученных выражений, — действительные числа).

Чтобы решить вторую часть задачи, заметим, что

$$(1+i)^4 = 1 + 4i + 6i^2 + 4i^3 + i^4 = -4 = (1-i)^4.$$

Следовательно,

$$j(4k+3) = (-4)^k j(3) = \frac{(-4)^k}{2} [(1+2i+i^2) + 1 - 2i + i^2] = 0.$$

Таким образом,

$$I(4k+3) = 0.$$

Упражнение 5.5-1. Пусть

$$I_k(\Phi) = \int_0^\pi \frac{\cos k\theta - \cos k\Phi}{\cos \theta - \cos \Phi} d\theta.$$

Показать, что для k , равного целому числу,

$$I_{k+2}(\Phi) - (2 \cos \Phi) I_{k+1}(\Phi) + I_k(\Phi) = 0, \quad I_0 = 0, \quad I_1 = \pi$$

и, следовательно,

$$I_k(\Phi) = \frac{\pi \sin k\Phi}{\sin \Phi}.$$

ГЛАВА 6

КОНЕЧНЫЕ РЯДЫ ФУРЬЕ

§ 6.1. Введение

Ряды Фурье

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx), \quad (6.1-1)$$

где

$$a_k = \frac{1}{\pi} \int_0^{2\pi} f(x) \cos kx \, dx, \quad b_k = \frac{1}{\pi} \int_0^{2\pi} f(x) \sin kx \, dx, \quad (6.1-2)$$

сыграли важную роль в развитии математики*). Наряду с обычной теорией, в которой функция предполагается известной на всем интервале длины 2π , существует теория, рассматривающая функции, заданные на дискретном множестве равноотстоящих точек. Мы будем полагать, что имеется $2N$ равноотстоящих точек, и оставим случай нечетного числа точек $2N+1$ в качестве упражнения для читателя.

Разложение в ряд Фурье основано на ортогональности функций 1 , $\cos kx$, $\sin kx$ по отношению к интегрированию по отрезку $[0, 2\pi]$

$$\begin{aligned} \int_0^{2\pi} \cos kx \cos mx \, dx &= \begin{cases} 0, & k \neq m, \\ \pi, & k = m \neq 0, \\ 2\pi, & k = m = 0, \end{cases} \\ \int_0^{2\pi} \sin kx \sin mx \, dx &= \begin{cases} 0, & k \neq m, \\ \pi, & k = m \neq 0, \end{cases} \\ \int_0^{2\pi} \sin kx \cos mx \, dx &= 0. \end{aligned} \quad (6.1-3)$$

§ 6.2. Ортогональность на дискретном множестве точек

Примечательно, что если вместо интегрирования мы используем суммирование, то функции

$$1, \cos x, \cos 2x, \dots, \cos (N-1)x, \cos Nx, \\ \sin x, \sin 2x, \dots, \sin (N-1)x$$

оказываются ортогональными на дискретном множестве точек

$$0, \quad \frac{\pi}{N}, \frac{2\pi}{N}, \dots, \frac{(2N-1)\pi}{N}. \quad (6.2-1)$$

Преобразование

$$x \rightarrow \frac{\pi}{N} x$$

приводит нас к системе функций:

$$1, \cos \frac{\pi}{N} x, \cos \frac{2\pi}{N} x, \dots, \cos \frac{(N-1)\pi}{N} x, \cos \frac{\pi}{N} Nx, \\ \sin \frac{\pi}{N} x, \sin \frac{2\pi}{N} x, \dots, \sin \frac{(N-1)\pi}{N} x. \quad (6.2-2)$$

*) Элементарное изложение рядов Фурье см. [17].

Соотношения ортогональности между ними ($k \leq N$, $m \leq N$)

$$\begin{aligned} \sum_{x=0}^{2N-1} \sin \frac{\pi}{N} kx \sin \frac{\pi}{N} mx &= \begin{cases} 0, & \text{если } k \neq m, \\ N, & \text{если } k = m \neq 0, \end{cases} \\ \sum_{x=0}^{2N-1} \sin \frac{\pi}{N} kx \cos \frac{\pi}{N} mx &= 0, \\ \sum_{x=0}^{2N-1} \cos \frac{\pi}{N} kx \cos \frac{\pi}{N} mx &= \begin{cases} 0, & \text{если } k \neq m, \\ N, & \text{если } k = m \neq 0, N, \dots, \\ 2N, & \text{если } k = m = 0, N, \dots, \end{cases} \end{aligned} \quad (6.2-3)$$

были доказаны в упражнении (3.2-3). Они могут рассматриваться как аналоги условия перпендикулярности (ортогональности) в обычном трехмерном пространстве. (См. § 17.6.)

Если функция может быть записана в виде

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{N-1} \left(a_k \cos \frac{\pi}{N} kx + b_k \sin \frac{\pi}{N} kx \right) + \frac{a_N}{2} \cos \pi x, \quad (6.2-4)$$

то можно использовать соотношения ортогональности для определения коэффициентов разложения. Чтобы получить a_k ($k = 1, \dots, N-1$), умножим обе части равенства на $\cos \frac{\pi}{N} mx$ и просуммируем по всем x .

Получим

$$\sum_{x=0}^{2N-1} f(x) \cos \frac{\pi}{N} mx = Na_m \quad (1 \leq m \leq N-1), \quad (6.2-5)$$

так как по (6.2-3) все остальные члены равны нулю. Заменяя $\cos \frac{\pi}{N} mx$ на $\sin \frac{\pi}{N} mx$, получаем

$$\sum_{x=0}^{2N-1} f(x) \sin \frac{\pi}{N} mx = Nb_m \quad (1 \leq m \leq N-1). \quad (6.2-6)$$

Окончательно имеем

$$\sum_{x=0}^{2N-1} f(x) = 2N \left(\frac{a_0}{2} \right) = Na_0, \quad \sum_{x=0}^{2N-1} f(x) \cos \pi x = 2N \left(\frac{a_N}{2} \right) = Na_N. \quad (6.2-7)$$

Последние две формулы имеют тот же вид, что (6.2-5) и показывают, почему выгоднее писать разложение с $a_0/2$ и $a_N/2$, а не с a_0 и a_N .

Упражнение 6.2-1. Рассмотреть разложение для нечетного числа $2N+1$ узловых точек.

$$\text{О т в е т: } f(x) = \frac{a_0}{2} + \sum_{k=1}^N \left(a_k \cos \frac{2\pi k}{2N+1} x + b_k \sin \frac{2\pi k}{2N+1} x \right).$$

§ 6.3. Точность разложения

В рассматриваемом разложении $2N$ значений $f(x)$ ($x=0, 1, \dots, 2N-1$) определяют $2N$ коэффициентов a_k, b_k и естественно ожидать, что сумма ряда в данных точках будет в точности равна значениям начальной функции (т. е. что система функций является полной по отношению к группе используемых точек). Чтобы доказать это предположение, зафиксируем $x = \bar{x}$ и перегруппируем сумму, применяя равенства (6.2-5)–(6.2-7), чтобы исключить a_k и b_k :

$$\begin{aligned} f(\bar{x}) &= \frac{1}{N} \sum_x f(x) \left[\frac{1}{2} + \sum_{k=1}^{N-1} \left(\cos \frac{\pi}{N} kx \cos \frac{\pi}{N} k\bar{x} + \sin \frac{\pi}{N} kx \sin \frac{\pi}{N} k\bar{x} \right) + \right. \\ &\quad \left. + \frac{1}{2} \cos \pi x \cos \pi \bar{x} \right] = \\ &= \frac{1}{N} \sum_x f(x) \left[\frac{1}{2} + \sum_{k=1}^{N-1} \cos \frac{\pi}{N} k(x - \bar{x}) + \frac{1}{2} \cos \pi x \cos \pi \bar{x} \right]. \end{aligned}$$

Теперь напомним

$$\sum_{k=1}^{N-1} \cos \frac{\pi}{N} k(x - \bar{x}) = \frac{1}{2} \sum_{k=1}^{N-1} \cos \frac{\pi}{N} k(x - \bar{x}) + \frac{1}{2} \sum_{k=1}^{N-1} \cos \frac{\pi}{N} k(x - \bar{x})$$

и во второй сумме используем тот факт, что

$$\cos \frac{\pi}{N} k(x - \bar{x}) = \cos \frac{\pi}{N} (2N - k)(x - \bar{x}),$$

чтобы получить

$$\sum_{k=1}^{N-1} \cos \frac{\pi}{N} k(x - \bar{x}) = \frac{1}{2} \sum_{k=1}^{N-1} \cos \frac{\pi}{N} k(x - \bar{x}) + \frac{1}{2} \sum_{k=N+1}^{2N-1} \cos \frac{\pi}{N} k(x - \bar{x}).$$

Используя это, имеем

$$f(\bar{x}) = \frac{1}{2N} \sum_x f(x) \left[\sum_{k=0}^{2N-1} \cos \frac{\pi}{N} k(x - \bar{x}) \right].$$

Но выражение в квадратных скобках равно

$$\sum_{k=0}^{2N-1} \cos \frac{\pi}{N} k (x - \bar{x}) = \begin{cases} 2N, & \text{если } x = \bar{x}, \\ 0, & \text{если } x \neq \bar{x}, \end{cases}$$

так что от суммы остается лишь член $x = \bar{x}$ и

$$f(\bar{x}) = \frac{1}{2N} f(\bar{x}) 2N = f(\bar{x}).$$

Таким образом, обе части уравнения дают одно и то же число.

Существует хорошо известное соотношение между суммой квадратов коэффициентов и суммой квадратов значений функции

$$\sum_x [f(x)]^2 = N \left[\frac{a_0^2}{2} + \sum_{k=1}^{N-1} (a_k^2 + b_k^2) + \frac{a_N^2}{2} \right]. \quad (6.3-1)$$

Чтобы доказать его, образуем сумму квадратов значений функции

$$\sum_x [f(x)]^2 = \sum_x \left[\frac{a_0}{2} + \sum_{k=1}^{N-1} \left(a_k \cos \frac{\pi}{N} kx + b_k \sin \frac{\pi}{N} kx \right) + \frac{a_N}{2} \cos \pi x \right]^2$$

и выполним умножение. Соотношения ортогональности (6.2-5)–(6.2-7) дают

$$\sum_x [f(x)]^2 = \left(\frac{a_0}{2} \right)^2 \cdot 2N + \sum_k (a_k^2 \cdot N + b_k^2 \cdot N) + \left(\frac{a_N}{2} \right)^2 \cdot 2N,$$

что совпадает с (6.3-1).

Этот результат может быть использован для вычисления суммы квадратов ошибок, когда используются члены лишь до $k = M < N-1$. Пусть $f_M(x)$ — сумма первых M гармоник, так что

$$f_M(x) = \frac{a_0}{2} + \sum_{k=1}^M \left(a_k \cos \frac{\pi}{N} kx + b_k \sin \frac{\pi}{N} kx \right).$$

Тогда

$$\begin{aligned} \sum_x (f - f_M)^2 &= \sum_x \left[\sum_{k=M+1}^{N-1} \left(a_k \cos \frac{\pi}{N} kx + b_k \sin \frac{\pi}{N} kx \right) + \frac{a_N}{2} \cos \pi x \right]^2 = \\ &= N \left[\sum_{k=M+1}^{N-1} (a_k^2 + b_k^2) + \frac{a_N^2}{2} \right] = \sum_x f^2 - N \left[\frac{a_0^2}{2} + \sum_{k=1}^M (a_k^2 + b_k^2) \right]. \end{aligned} \quad (6.3-2)$$

Другими словами, использование членов лишь до $k = M$ в разложении функции (6.2-4) дает нам приближение к $f(x)$, ошибка которого в смысле наименьших квадратов определяется равенством (6.3-2). Исследуя, как уменьшается (6.3-2) при увеличении M , мы можем оценить значение взятых дополнительно членов в разложении Фурье.

§ 6.4. Вычисление коэффициентов

Коэффициенты разложения (6.2-4) даются формулами

$$\left. \begin{aligned} a_k &= \frac{1}{N} \sum_{x=0}^{2N-1} f(x) \cos \frac{\pi}{N} kx \quad (k=0, 1, 2, \dots, N), \\ b_k &= \frac{1}{N} \sum_{x=0}^{2N-1} f(x) \sin \frac{\pi}{N} kx \quad (k=1, 2, \dots, N-1). \end{aligned} \right\} \quad (6.4-1)$$

Показано, что эти коэффициенты можно легко вычислить, зная лишь $\cos \frac{\pi}{N} k$ и $\sin \frac{\pi}{N} k$, без непосредственного вычисления других значений синуса и косинуса. Это делается следующим образом: пусть

$$\left. \begin{aligned} U_0 &= 0, \quad U_1 = f(2N-1), \\ U_m &= \left(2 \cos \frac{\pi}{N} k \right) U_{m-1} - U_{m-2} + f(2N-m) \\ &\quad (m=2, 3, \dots, 2N-1). \end{aligned} \right\} \quad (6.4-2)$$

Тогда

$$\left. \begin{aligned} Na_k &= \sum_{x=0}^{2N-1} f(x) \cos \frac{\pi}{N} kx = \left(\cos \frac{\pi}{N} k \right) U_{2N-1} - U_{2N-2} + f(0), \\ Nb_k &= \sum_{x=0}^{2N-1} f(x) \sin \frac{\pi}{N} kx = \left(\sin \frac{\pi}{N} k \right) U_{2N-1}. \end{aligned} \right\} \quad (6.4-3)$$

Прямое формальное доказательство не показывает механизма работы этого метода, так что мы дадим несколько более длинное доказательство, которое к тому же даст еще и дополнительный результат.

Начнем с определения

$$\left. \begin{aligned} V_0 &= 0, \quad V_1 = 1, \\ V_m &= (2 \cos t) V_{m-1} - V_{m-2} \quad (m=2, 3, \dots) \end{aligned} \right\} \quad (6.4-4)$$

и покажем сначала, что

$$V_m = \frac{\sin mt}{\sin t}. \quad (6.4-5)$$

Так как (6.4-5) верно для $m=0; 1$, то достаточно проверить, что

$$\begin{aligned} (2 \cos t) V_{m-1} - V_{m-2} &= \frac{2 \cos t \sin (m-1) t}{\sin t} - \frac{\sin (m-2) t}{\sin t} = \\ &= \frac{\sin mt + \sin (m-2) t - \sin (m-2) t}{\sin t} = \frac{\sin mt}{\sin t} = V_m. \end{aligned}$$

Мы замечаем также, что

$$\begin{aligned}\cos t(V_m) - V_{m-1} &= \frac{\cos t \sin mt - \sin(m-1)t}{\sin t} = \\ &= \frac{\frac{1}{2} [\sin(m+1)t + \sin(m-1)t] - \sin(m-1)t}{\sin t} = \\ &= \frac{\sin(m+1)t - \sin(m-1)t}{2 \sin t} = \cos mt. \quad (6.4-6)\end{aligned}$$

Таким образом, зная значения $\sin t$ и $\cos t$, мы можем вычислить, применяя (6.4-4)–(6.4-6), значения $\sin 2t$, $\cos 2t$, $\sin 3t$, $\cos 3t$, ..., $\sin kt$, $\cos kt$. При этом мы делаем не более, чем

$$\left\{ \begin{array}{l} 3k \text{ умножений} \\ 2k \text{ вычитаний} \end{array} \right\} \quad \text{или} \quad \left\{ \begin{array}{l} 2k \text{ умножений} \\ 3k \text{ вычитаний} \end{array} \right\}.$$

При вычислении значений этим методом происходит, конечно, постепенная потеря точности, но обычно она незначительна.

Покажем теперь, как применить равенства (6.4-4)–(6.4-6) к доказательству (6.4-3), используя (6.4-2).

Во-первых, рассмотрим случай, когда все $f(x) = 0$, кроме $f(j)$, и положим $t = \frac{\pi}{N} k$. Из (6.4-2) видно сразу, что

$$U_0 = U_1 = \dots = U_{2N-j-1} = 0, \quad U_{2N-j} = f(j).$$

Таким образом, U_{2N-j} соответствует V_1 , за исключением того, что все U , которые вычисляются, будут в $f(j)$ раз больше, чем соответствующие V . Когда мы дойдем до U_{2N-1} , мы сделаем ровно j шагов и ему будет соответствовать V_j . По (6.4-5) и (6.4-6) результаты, которые мы нашли в (6.4-3), равны

$$f(j) \sin j \frac{\pi}{N} k, \quad f(j) \cos j \frac{\pi}{N} k.$$

Вернемся теперь к произвольной функции $f(x)$. Она может быть представлена в виде суммы функций только что рассмотренного типа. В силу линейности (6.4-2) и (6.4-3) отдельные результаты для каждой из компонент можно сложить, что и доказывает (6.4-3).

Для нахождения каждой пары a_k, b_k требуется около $6N$ арифметических операций, так что для всех $2N$ коэффициентов нужно сделать около $6N^2$ действий. Если бы матрица синусов и косинусов была в памяти машины и для получения коэффициентов мы просто умножали бы матрицу на вектор, то при этом пришлось бы затратить $8N^2$ действий. Таким образом, можно сказать, что предложенный метод экономит машинное время даже по сравнению с очень благоприятной ситуацией, когда дана матрица синусов и косинусов.

Из двух значений $\sin \frac{\pi}{N}$, $\cos \frac{\pi}{N}$, используя метод V -обозначений, можно получить

$$\sin \frac{\pi}{N} k, \quad \cos \frac{\pi}{N} k \quad (k = 2, 3, \dots, N-1).$$

Основной процесс в V -обозначении теперь дает требуемые коэффициенты. При соответствующем переплетении V - и U -процессов в каждый момент необходимо лишь $2N$ ячеек для текущих данных и несколько счетчиков. Программы для вычисления V и U также довольно короткие, а следовательно, не занимают много места в памяти.

§ 6.5. Метод двенадцати ординат

Особый случай, когда число точек равно 12 ($N=6$), представляет интерес, так как часто встречается и легко выполняется вручную. Мы имеем

$$6a_k = \sum_{x=0}^{11} f(x) \cos \frac{\pi}{6} kx, \quad 6b_k = \sum_{x=0}^{11} f(x) \sin \frac{\pi}{6} kx.$$

Разбивая каждую сумму по x на две: от 0 до 6 и от 7 до 11, и положив во второй части $x=12-x'$, получим

$$6a_k = \sum_{x=0}^6 f(x) \cos \frac{\pi}{6} kx + \sum_{x'=1}^5 f(12-x') \cos \frac{\pi}{6} kx',$$

$$6b_k = \sum_{x=0}^6 f(x) \sin \frac{\pi}{6} kx - \sum_{x'=1}^5 f(12-x') \sin \frac{\pi}{6} kx'.$$

	$f(0)$	$f(1)$	$f(2)$	$f(3)$	$f(4)$	$f(5)$	$f(6)$
Сумма	$s(0)$	$s(1)$	$s(2)$	$s(3)$	$s(4)$	$s(5)$	$s(6)$
Разность		$t(1)$	$t(2)$	$t(3)$	$t(4)$	$t(5)$	

Тогда имеем

$$6a_k = \sum_{x=0}^6 s(x) \cos \frac{\pi}{6} kx, \quad 6b_k = \sum_{x=1}^5 t(x) \sin \frac{\pi}{6} kx.$$

Разобьем опять отрезок на две части, на этот раз от 0 до 3 и от 4 до 6, положив $x=6-x'$ во втором отрезке. Тогда

$$6a_k = \sum_{x=0}^3 s(x) \cos \frac{\pi}{6} kx + (-1)^k \sum_{x'=0}^2 s(6-x') \cos \frac{\pi}{6} kx',$$

$$6b_k = \sum_{x=1}^3 t(x) \sin \frac{\pi}{6} kx - (-1)^k \sum_{x'=1}^2 t(6-x') \sin \frac{\pi}{6} kx'.$$

Теперь можно написать

	$s(0)$	$s(1)$	$s(2)$	$s(3)$	$t(1)$	$t(2)$	$t(3)$
	$\frac{s(6)}{u(0)}$	$\frac{s(5)}{u(1)}$	$\frac{s(4)}{u(2)}$	$\frac{t(5)}{p(1)}$	$\frac{t(4)}{p(2)}$	$\frac{t(3)}{p(3)}$	
Сумма	$u(0)$	$u(1)$	$u(2)$	$u(3)$	$p(1)$	$p(2)$	$p(3)$
Разность	$v(0)$	$v(1)$	$v(2)$		$q(1)$	$q(2)$	

Результат, выписанный полностью, выглядит так:

$$6a_0 = u(0) + u(1) + u(2) + u(3) = [u(0) + u(3)] + [u(1) + u(2)],$$

$$6a_1 = v(0) + \frac{\sqrt{3}}{2}v(1) + \frac{1}{2}v(2) = \left[v(0) + \frac{1}{2}v(2)\right] + \frac{\sqrt{3}}{2}v(1),$$

$$6a_2 = u(0) + \frac{1}{2}[u(1) - u(2)] - u(3) = [u(0) - u(3)] +$$

$$+ \frac{1}{2}[u(1) - u(2)],$$

$$6a_3 = v(0) - v(2),$$

$$6a_4 = u(0) - \frac{1}{2}u(1) - \frac{1}{2}u(2) + u(3) = [u(0) + u(3)] - \frac{1}{2}[u(1) + u(2)],$$

$$6a_5 = v(0) - \frac{\sqrt{3}}{2}v(1) + \frac{1}{2}v(2) = \left[v(0) + \frac{1}{2}v(2)\right] - \frac{\sqrt{3}}{2}v(1),$$

$$6a_6 = u(0) - [u(1) - u(2)] - u(3) = [u(0) - u(3)] - [u(1) - u(2)],$$

$$6b_1 = \frac{1}{2}p(1) + \frac{\sqrt{3}}{2}p(2) + p(3) = \left[\frac{1}{2}p(1) + p(3)\right] + \frac{\sqrt{3}}{2}p(2),$$

$$6b_2 = \frac{\sqrt{3}}{2}[q(1) + q(2)],$$

$$6b_3 = p(1) - p(3),$$

$$6b_4 = \frac{\sqrt{3}}{2}[q(1) - q(2)],$$

$$6b_5 = \frac{1}{2}p(1) - \frac{\sqrt{3}}{2}p(2) + p(3) = \left[\frac{1}{2}p(1) + p(3)\right] - \frac{\sqrt{3}}{2}p(2).$$

Здесь приходится делать меньше 60 арифметических операций и большинство из них — простые сложения. Это можно сравнить с $6N^3 = 6^3 = 216$ операциями, необходимыми в методе § 6.4. Программа, конечно, несколько длиннее, но писать ее очень легко, так как нет сложной логики и требуется лишь одно «непростое» число

$$\frac{\sqrt{3}}{2} = 0,866\ 025\ 407\ 5.$$

Общие методы могут и должны применяться в случаях, отличных от $N=6$.

Упражнение 6.5-1. Найти вручную разложение Фурье для $f(x) = x(12-x)$ ($x=0, 1, \dots, 11$).

§ 6.6. Методы с минимумом умножений

Вообще в задачи этой книги не входит вдаваться в детали того, как организовывать вычисление, чтобы экономить машинное время. Однако показать, как можно эффективно вычислять ряды Фурье, необходимо, так как обычно предпочитают многочленные разложения, а также потому, что при использовании сумм Фурье обычно тратится много лишнего машинного времени.

Число умножений можно свести к минимуму, выполняя сначала все сложения и максимально используя приведение тригонометрических функций к значениям в первой четверти. Вот почему работает метод § 6.5. Всякий раз, когда порядок гармоник является делителем числа $2N$, некоторые из возможных значений тригонометрической функции не требуются. Поэтому на практике часто используются значения $2N$, равные 12, 24 и 60.

§ 6.7. Разложение по косинусам

Если допустить, что функция $f(x)$ периодическая с периодом $2N$, то при использовании значений $x=0, 1, \dots, 2N-1$ или $-(N-1), -(N-2), \dots, 0, 1, \dots, N$ результат будет один и тот же. Теперь предположим, что задана функция, определенная для $x=0, 1, \dots, N$, и мы хотим разложить ее в ряд только по косинусам. Определим $f(-x) = f(x)$ для $x=1, 2, \dots, N-1$. Теперь $f(x)$ — четная функция. Когда мы находим разложение Фурье, то получаем, что все $b_k = 0$, так как

$$b_k = \frac{1}{N} \sum_{x=-N+1}^N f(x) \sin \frac{\pi k}{N} x,$$

а синус — нечетная функция. Таким образом получаем разложение по косинусам. Формулы для a_k можно упростить:

$$\begin{aligned} a_k &= \frac{1}{N} \sum_{x=-N+1}^{N-1} f(x) \cos \frac{\pi k}{N} x = \\ &= \frac{2}{N} \left[\frac{1}{2} f_0 + \sum_{x=1}^{N-1} f(x) \cos \frac{\pi k}{N} x + \frac{1}{2} (-1)^k f_N \right]. \end{aligned}$$

Определяя $f(x)$ как нечетную функцию, можно аналогичным образом получить разложение по синусам.

Заметим также, что если мы имеем область $0 \leq x \leq R$, то замена

$$x = \frac{R}{2} (1 + \cos t)$$

делает функцию четной периодической функцией t . Эта замена дает ряд преимуществ в вопросах сходимости, которые будут обсуждаться в гл. 22.

Упражнения

6.7-1. Функция $f(x)$ определена на интервале $0 \leq x \leq \pi/2$. Показать, что можно продолжить функцию на $-\pi < x \leq \pi$ так, чтобы встречались лишь нечетные (четные) гармоники косинусов.

6.7-2. Рассмотреть в деталях задачу нахождения рядов Фурье только по синусам. По четным (или нечетным) гармоникам синусов.

§ 6.8. Локальные ряды Фурье

Во многих задачах явление меняется медленно, и идея «локальных рядов Фурье», коэффициенты которых медленно меняются с течением времени, совершенно естественна. Также более естественно центрировать область около начала и использовать значения от $-N+1$ до N , чем от 0 до $2N-1$. Таким образом, из равенств (6.2-5)–(6.2-7) имеем

$$\left. \begin{aligned} a_m(t) &= \frac{1}{N} \sum_{x=-N+1}^N f(x+t) \cos \frac{\pi}{N} mx, \\ m &= 0, 1, \dots, N, \\ b_m(t) &= \frac{1}{N} \sum_{x=-N+1}^N f(x+t) \sin \frac{\pi}{N} mx, \\ m &= 1, 2, \dots, N-1. \end{aligned} \right\} \quad (6.8-1)$$

Легко получить $a(t+1)$ и $b(t+1)$ из $a(t)$ и $b(t)$. Например,

$$a_m(t+1) = \frac{1}{N} \sum_{x=-N+1}^N f(x+t+1) \cos \frac{\pi}{N} m x.$$

Положим $x+1 = x'$, тогда

$$\begin{aligned} a_m(t+1) &= \frac{1}{N} \sum_{x'=-N}^{N+1} f(x'+t) \cos \frac{\pi}{N} m (x'-1) = \\ &= \frac{1}{N} \left\{ \sum_{x'=-N+1}^N f(x'+t) \cos \frac{\pi}{N} m (x'-1) + \right. \\ &\quad \left. + (-1)^m [f(N+1+t) - f(-N+t)] \right\}. \end{aligned}$$

Раскрывая косинус, получаем

$$\begin{aligned} a_m(t+1) &= \left[a_m(t) \cos \frac{\pi}{N} m + b_m(t) \sin \frac{\pi}{N} m \right] + \\ &\quad + \frac{(-1)^m}{N} [f(N+1+t) - f(-N+t)]. \end{aligned}$$

Заметим, что мы пользуемся фиксированными (для данной частоты) множителями $\cos \frac{\pi m}{N}$ и $\sin \frac{\pi m}{N}$ и складываем или вычитаем два новых значения функции. Аналогично для синусов:

$$b_m(t+1) = \left[-a_m(t) \sin \frac{\pi}{N} m + b_m(t) \cos \frac{\pi}{N} m \right].$$

ПРИБЛИЖЕНИЕ МНОГОЧЛЕНАМИ — КЛАССИЧЕСКИЙ ЧИСЛЕННЫЙ АНАЛИЗ

ГЛАВА 7

ВВЕДЕНИЕ В МНОГОЧЛЕННЫЕ ПРИБЛИЖЕНИЯ

§ 7.1. Ориентация

Первые работы любого вычислителя обычно связаны с вычислением значений функции. Главными инструментами здесь являются «общее чутье» и «маленькие хитрости». Цель этой книги — показать главные идеи в области вычислений; мы будем лишь изредка иллюстрировать аспект «маленьких хитростей», хотя они до сих пор являются важной частью искусства вычисления.

Второе, с чем обычно сталкиваются, — это интерполяция недостающих в таблице значений, например в логарифмической или тригонометрической таблице. Вообще говоря, при интерполяции нам дано несколько узлов функции и нужно вычислить приближенно некоторые значения, которых нет в таблице. В большинстве таблиц сделано предположение, что функция ведет себя между последовательно взятыми точками, как прямая, хотя иногда предполагается, что она ведет себя, как квадратный трехчлен и даже многочлен более высокой степени.

Часто думают, что главные проблемы численного анализа сосредоточены вокруг интерполяции; но это не так. К ним относятся скорее такие операции, как интегрирование, дифференцирование, нахождение нулей, максимизация и т. д. в случаях, когда все, что мы имеем или можем вычислить, — это некоторые узлы функции, причем и они обычно известны не точно, а приближенно, так как бывают испорчены погрешностью округления.

Классический численно-аналитический подход заключается в том, чтобы использовать некоторые узлы функции для получения приближающего многочлена и затем выполнить аналитическую операцию над этим многочленом. Этот процесс может быть назван «аналитической заменой» ([38], стр. 51), так как функция, которую невозможно обработать, заменяется другой функцией, над которой уже можно выполнить аналитическую операцию. Например, в способе Ньютона для

нахождения нуля функции $y = f(x)$ дается приближенное значение x_1 и вместо кривой используется прямая

$$y - y_1 = y'_1(x - x_1),$$

которая касается графика функции в точке (x_1, y_1) . Подставляя $y = 0$, получаем значение x , являющееся корнем этой новой функции,

$$x = x_1 - \frac{y_1}{y'_1}.$$

Это новое значение x используется как следующее приближенное значение корня.

Поскольку с многочленами легко обращаться, большая часть классического численного анализа основывается на приближении многочленами. Однако для многих целей предпочитают другие классы функций; они будут изучаться после того, как будет рассмотрена аппроксимация многочленами.

Выбрав узлы и класс приближающих функций, мы должны еще выбрать одну определенную функцию из этого класса посредством некоторого критерия — некоторой меры приближения или «согласия». Самый широко применяемый критерий состоит в требовании того, чтобы приближающая функция совпадала с заданными значениями в узловых точках. Другой более общий критерий — «наименьшие квадраты» — означает, что «сумма квадратов отклонений между данными узлами и приближающей функцией в узловых точках должна быть минимальной». Однако иногда применяются и другие критерии.

Прежде чем начать вычисления, мы должны решить также, какую точность мы хотим иметь в ответе и какой критерий мы изберем для измерения этой точности. Все изложенное можно сформулировать в виде четырех вопросов:

1. *Какие узлы мы будем использовать?*
2. *Какой класс приближающих функций мы будем использовать?*
3. *Какой критерий согласия мы применим?*
4. *Какую точность мы хотим?*

Упражнение 7.1-1. Взяв три члена ряда Тейлора

$$y = y(x_1) + (x - x_1)y'(x_1) + \frac{(x - x_1)^2 y''(x_1)}{2},$$

приблизить функцию в точке x_1 , обобщив формулы Ньютона для нахождения корней функции.

$$\text{Ответ: } x = x_1 - \frac{y(x_1)}{y'(x_1)} \left[1 + \frac{y(x_1) y''(x_1)}{2y'^2(x_1)} \right].$$

§ 7.2. Альтернативные формулировки

В описанном только что процессе, найдя приближающую функцию, мы затем применяем к ней аналитическую операцию и вычисляем результат в некоторой точке или точках. Обычно окончательный результат оказывается линейной комбинацией значений в первоначальных узлах. Это подсказывает, что можно прийти непосредственно от узловых точек к ответу, не получая по дороге аппроксимирующей функции.

Чтобы сделать эти замечания более определенными, предположим, что имеется некоторый линейный оператор L , действующий в некотором классе функций, так что

$$L[af(x) + bg(x)] = aL[f(x)] + bL[g(x)]. \quad (7.2-1)$$

Интегрирование, дифференцирование и интерполяция — типичные примеры линейных операторов. Нахождение нуля — не линейный оператор.

Результат $L[f(x)]$ будем искать, исходя из данных в узлах значений функции и ее производных. Таким образом, мы ищем формулу вида

$$\begin{aligned} L[f(x)] = & a_1 f(x_1) + a_2 f(x_2) + \dots + a_n f(x_n) + \\ & + b_1 f'(x_1) + b_2 f'(x_2) + \dots + b_n f'(x_n) + \dots + k_1 f^{(k-1)}(x_1) + \\ & + k_2 f^{(k-1)}(x_2) + \dots + k_n f^{(k-1)}(x_n), \end{aligned} \quad (7.2-2)$$

где многие коэффициенты могут быть равны нулю. Для простоты представим, что у нас есть лишь значения функции, а все коэффициенты b, c, \dots, k равны нулю.

Предположим, что используется приближение многочленом и требуется точное совпадение его с данными значениями в узловых точках. Возможны три эквивалентных способа определения коэффициентов a_i .

Первый способ, который уже рассматривался, состоит в нахождении приближающего многочлена и применении к нему оператора L ; это — так называемый метод аналитической замены.

Второй способ состоит в том, чтобы разложить $f(x)$ в ряд Тейлора, подставить разложение в левую и правую части и приравнять коэффициенты при одинаковых производных. Поясним его построением, например, формулы Симпсона

$$\int_{-1}^1 f(x) dx = a_{-1} f(-1) + a_0 f(0) + a_1 f(1). \quad (7.2-3)$$

Так как

$$f(x) = f(0) + xf'(0) + \frac{x^2}{2!} f''(0) + \frac{x^3}{3!} f'''(0) + \frac{x^4}{4!} f^{IV}(0) + \dots,$$

то в левой части равенства будем иметь

$$\int_{-1}^1 f(x) dx = 2f(0) + \frac{2}{3!} f''(0) + \frac{2}{5!} f^{IV}(0) + \dots,$$

тогда как справа

$$a_{-1}f(-1) = a_{-1}f(0) - a_{-1}f'(0) + \frac{a_{-1}}{2!} f''(0) - \\ - \frac{a_{-1}}{3!} f'''(0) + \frac{a_{-1}}{4!} f^{IV}(0) + \dots,$$

$$a_0 f(0) = a_0 f(0),$$

$$a_1 f(1) = a_1 f(0) + a_1 f'(0) + \frac{a_1}{2!} f''(0) + \frac{a_1}{3!} f'''(0) + \frac{a_1}{4!} f^{IV}(0) + \dots$$

Приравнявая, пока это возможно, коэффициенты при $f(0)$, $f'(0)$, ..., получаем

$$2 = a_{-1} + a_0 + a_1,$$

$$0 = -a_{-1} + a_1,$$

$$\frac{2}{3} = a_{-1} + a_1,$$

$$0 = -a_{-1} + a_1,$$

$$\frac{2}{5} = a_{-1} + a_1 + E,$$

где E — ошибка. Решение этих уравнений дает значения $a_{-1} = a_1 = 1/3$, $a_0 = 4/3$, которые приводят к обычной формуле

$$\int_{-1}^1 f(x) dx = 1/3 f_{-1} + 4/3 f_0 + 1/3 f_1 \quad (7.2-4)$$

Третий способ требует, чтобы формула была точной для последовательности функций $f(x) = 1, x, x^2, \dots$ до возможно более высокой степени. Он полезен для получения специальных формул, и мы будем использовать его для этой цели в большинстве случаев. Однако при решении конкретной задачи метод аналитической замены, при котором в первую очередь строится приближающая функция для исходных данных, часто бывает лучше, так как он дает возможность получающему результаты видеть своими глазами, какая именно аппроксимация делается. Так, в примере § 1.9, когда находился аппроксимирующий многочлен, мы вычисляли его в точках посередине между узловыми и строили получившиеся значения, чтобы видеть, какой график они дают.

Покажем, что эти три способа эквивалентны и дают один и тот же ответ. Третий способ, который делает формулу точной для различных степеней x до некоторой степени, скажем n , делает формулу точной

также и для любого многочлена степени, не превосходящей n . Таким образом, если функция есть многочлен степени n (или меньшей), то формула точная. Но если функция есть многочлен, то ее ряд Тейлора оканчивается членом n -го порядка и второй способ даст тот же результат. Обратное также верно, следовательно, второй способ эквивалентен третьему.

Чтобы показать, что первый способ, который заключается в нахождении интерполяционного многочлена и оперировании с ним, эквивалентен двум другим, необходимо иметь в распоряжении некоторую теорию интерполяционных многочленов, а ее мы будем развивать лишь в следующей главе. Пока предположим, что если даны значения f_i ($i = 1, 2, \dots, n$) функции, то можно построить многочлен*)

$$p(x) = L_1(x)f_1 + L_2(x)f_2 + \dots + L_n(x)f_n \quad (7.2-5)$$

который для $f(x) = 1, x, x^2, \dots, x^{n-1}$ совпадает с $f(x)$. Применив оператор $L[f(x)]$ к левой и правой частям (7.2-5), получим формулу типа (7.2-2), которая является точной для функций $1, x, x^2, \dots, x^{n-1}$.

Обратно, так как $L[f(x)]$ линеен, то формула, точная для функций $1, x, x^2, \dots, x^{n-1}$, точна и для любой их линейной комбинации, следовательно, для линейной комбинации (7.2-5). Таким образом, третий способ эквивалентен первому.

Третий способ, которым мы будем пользоваться для нахождения формул, является наиболее употребительным. Однако, чтобы найти формулу этим общим приемом, часто требуется больше работы, чем если применить различные трюки, которые приводятся в учебниках. В защиту нашего метода мы скажем: «Принимая во внимание огромное количество имеющихся теперь знаний, лучше применять один общий метод, до некоторой степени непроизводительный, чем изучать множество специальных трюков».

Окончательные формулы также безразличны к методу их получения. Метод, будучи общим, легко дает новые формулы и обеспечивает аппарат для дальнейшего исследования. В последнее время принятие такого единообразного метода направлено к постепенному перекладыванию всей работы на вычислительную машину.

Одна из особенностей этой книги — использование идей теории информации. Может быть, это просто предрассудок автора, но кажется разумным рассматривать универсальную вычислительную машину как «переработчик» информации. Многие из идей теории информации тесно связаны с функциями с ограниченным спектром, которые рассматриваются в третьей части этой книги. Хотя, как мы видели в главе 2, теория, которая рассматривает шум как основной элемент, еще недостаточно развита, некоторый сдвиг в этом направлении на-

*) Многочлены L_n определены в § 8.3; их не следует путать с линейным оператором L в равенстве (7.2-1).

мечен в новых книгах и публикациях. Лучшее, что мы можем сделать в настоящий момент, это запомнить, что почти каждое число, которое мы получаем, будет иметь некоторое «шумовое загрязнение». Поэтому необходимо исследовать формулы, которые здесь приводятся, в свете того, как они будут распространять шум.

Упражнение 7.2-1. Вывести правило трапеций первым, вторым и третьим способами.

§ 7.3. Узловые точки, информация

Первым из четырех вопросов, заданных в конце § 7.1, был вопрос «*какие узлы мы будем использовать*»? В принципе это — вопрос статистики, особенно области, называемой «планированием экспериментов». На практике узловые точки часто заданы внешними обстоятельствами или мы пользуемся множеством равноудаленных точек. В последнем случае мы лицом к лицу сталкиваемся с *теоремой выборки* из теории информации, которая приводится в третьей части, но на самом деле объясняет многое из материала второй части. Таким образом, в некотором отношении многое во второй части является несколько поверхностным.

Чтобы увидеть, что проблема выбора узлов реальна, представьте, что вас просят вычислить интеграл

$$\int_0^1 f(x) dx$$

для некоторой сложной функции, которую вы не можете проинтегрировать аналитически. Вас попросили в действительности оценить площадь области под кривой, основываясь на ряде узлов (скажем, их n), или, что то же самое, оценить среднее значение $f(x)$. Идеально, нужно знать о функции очень много, чтобы расположить узлы наилучшим образом. Допуская различные свойства функции, мы получаем различные способы расположения узлов (вследствие привычки автора слова «узлы» и «информация» часто будут применяться взаимозаменяемо).

Другим примером задачи, в которой возникает вопрос «*где лежит информация*?» — является задача нахождения производной в точке (хотя, в противоположность интегралу, производная — локальная характеристика). Здравый смысл диктует использование информации, близкой к точке, в которой надо оценить производную. С другой стороны, если узлы взять слишком близко, то шум в вычислении или измерении узловых точек будет мешать вычислению с хорошей точностью. Таким образом, соответствующий ответ на вопрос «где взять узлы?» требует оценки шума, который портит всю информацию. Как было замечено раньше, соответствующей теории шума нет, но развитие медленно идет в этом направлении.

§ 7.4. Класс функций

Второй из четырех вопросов был: *«какой класс аппроксимирующих функций мы будем применять?»*

Существуют три класса или группы функций, широко применяемых в численном анализе. Первая группа включает в себя линейные комбинации функций

$$1, x, x^2, \dots, x^n,$$

что совпадает с классом всех многочленов степени n (или меньше). Второй класс образуют функции

$$\cos a_i x, \quad \sin a_i x.$$

Этот класс имеет отношение к рядам Фурье и интегралу Фурье. Они рассматриваются в третьей части. Третья группа образуется функциями $e^{-a_i x}$. Эти функции часто встречаются в реальных ситуациях. К ним, например, приводят обычные задачи накопления и распада. Они также рассматриваются в третьей части (см. гл. 26).

Каждая из этих трех групп обладает одним важным свойством: конечное множество функций такой группы переходит само в себя, когда x заменяется на $x+k$. Например, если $P(x)$ — многочлен степени n , то $P(x+k)$ — также многочлен степени n по x , хотя, конечно, их коэффициенты различны. То же самое относится к двум другим группам.

Это свойство важно, так как оно подразумевает, что, когда мы выбираем множество аппроксимирующих функций, нам не требуется иметь сведений о начале отсчета. Использование любого другого конечного множества аппроксимирующих функций (кроме комбинации этих трех) подразумевает существование естественного начала отсчета, так как выбор начала отсчета будет влиять на результат*). Если в задаче есть особенность, это автоматически определяет естественное начало отсчета; обычно характер особенности подсказывает, какую группу аппроксимирующих функций надо использовать (см. третью часть, гл. 27). Но в большинстве задач нет естественного начала отсчета, и мы вынуждены выбирать одну из этих трех групп или комбинации их, если выбор не должен повлиять на ответ.

Множество линейных комбинаций $1, x, x^2, \dots, x^n$ имеет еще одно важное свойство, а именно: как множество оно также не изменится при замене x на kx . Другие две группы этим свойством не обладают. Таким образом, если в задаче нет естественного масштаба, то мы вынуждены использовать многочлены или воздействовать на

*) Мы не будем доказывать здесь это утверждение, но просто используем его, чтобы объяснить, почему рассматриваются в основном эти три группы.

ответ самим выбором масштаба. Но даже если множество аппроксимирующих функций обладает свойством независимости от масштаба, во многих задачах, особенно тех, в которых функции зависят от времени, есть естественный масштаб. Поэтому преобладание многочленов над другими двумя группами весьма прискорбно. Сравнительно запущенное состояние использования двух других групп означает, что теории их не были развиты так же далеко, как для многочленов, и в результате мы считаем эти теории более трудными, чем они есть.

Отношения двух многочленов, или, проще, рациональные функции, также обладают свойством независимости от масштаба и для них существует постепенно развивающаяся и вполне удовлетворительная теория (см. гл. 20). Мы обычно не сталкиваемся с константами, возведенными в многочленную степень, или многочленами, возведенными в многочленную степень; каждый из этих классов также обладал бы свойством независимости от масштаба.

§ 7.5. Согласие

Третьим из четырех вопросов был вопрос: *«каким критерием согласия мы будем пользоваться»?* Классическим является ответ: *«точное совпадение в узловых точках»*, который имеет преимущество простоты теории и выполнения вычислений, но также и неудобство из-за игнорирования шума. В основном вторая часть книги посвящена точному совпадению, но нужно помнить, что этот критерий пренебрегает особенно важными аспектами типичной вычислительной ситуации. Мы уже видели, что критерий точного совпадения в узловых точках в случае использования многочленов дает то же, что получение формулы, точной для $1, x, x^2, \dots, x^n$.

Другой относительно хороший критерий — это *«наименьшие квадраты»*. Он означает, что сумма квадратов отклонений в узловых точках должна быть сделана наименьшей возможной или, проще, минимизирована. Этот критерий использует избыточную информацию, чтобы получить некоторое сглаживание шума. Он популярен у математиков вследствие красоты математической теории, которая ему соответствует, и у физиков, потому что они полагают, что наименьшие квадраты — это принцип природы. Но человек, занимающийся вычислениями, имеет право смотреть на этот критерий с подозрением в каждом конкретном случае (см. главы 17 и 18).

Третий критерий связывается с именем Чебышева. Основная идея его состоит в том, чтобы уменьшить максимальное отклонение до минимума. Последний критерий становится особенно популярным в наши дни (см. гл. 19). Очевидно, возможны и другие критерии. Выбор какого-либо из них обычно определяется физической задачей, из которой возникла вычислительная. Мы попытаемся дать краткую характеристику различным критериям, когда будем вводить их, но

так как это — учебник по численным методам, то мы не станем тратить время на то, чтобы обучить читателя умению находить нужный критерий. Однако выбор критерия часто сильно влияет как на то, что должно быть сделано, так и на результат, который получится; поэтому этот вопрос заслуживает внимательного рассмотрения перед началом вычислений.

§ 7.6. Точность

Последним из четырех был вопрос: *«какую точность мы хотим получить»?* Внешний вид ответа, который обычно дается, таков: *«3 или 4 десятичных знака в ответе».*

Во многих задачах этого недостаточно: такой ответ игнорирует основную сторону вопроса. Рассмотрим, например, решение системы линейных алгебраических уравнений для неизвестных x_i . Нам нужно ответить на следующие вопросы:

1. Должны ли быть точными x_i ?
2. Должны ли быть маленькими остатки уравнений после того, как вычисленные x_i подставлены в них?
3. Должна ли данная группа уравнений быть близкой к группе уравнений, для которой x_i суть точные ответы?

Очевидно, что в определенной конкретной ситуации может годиться еще какая-либо другая мера точности. Обычно отвечают утвердительно на первый вопрос; но лишь иногда такой ответ является самым подходящим.

Второй критерий часто является надлежащим, потому что стремятся обратить в нуль уравнения и вычислить x_i , которые дают это обращение в нуль.

Третий критерий также очень полезен и постепенно все больше и больше используется в приложениях, так же как и в формальном анализе ошибок. В известном смысле говорят: «Так как ни ваши измерения, ни ваша физическая теория не являются точными, то как близко к ним я должен подойти в своих вычислениях?» Во многих отношениях это — самый основной критерий, критерий, из которого делаются предположения для других критериев. Ошибки в известном весе снаряда или планеты или даже в величине светового давления могут превратиться в равноценные ошибки, допускаемые на каждом шагу вычисления траектории, и учет этого обстоятельства имеет большее значение, чем задание произвольной «допустимой ошибки», которое будет достигнуто в конце траектории. Таким образом, мы видим, что путь, на котором мы отвечаем на четыре вопроса:

1. *Какие узловые точки мы выберем?*
2. *Какой класс аппроксимирующих функций мы выберем?*
3. *Какой критерий согласия мы выберем?*
4. *Какую точность мы хотим?*

может значительно влиять на ответы, которые получаются из вычисления. Также должно быть ясно, что ответы на эти вопросы должны быть найдены в первоначальной задаче, а не в математических трактатах или даже в книгах по численному анализу. *Здравая вычислительная практика требует постоянного исследования изучаемой задачи не только перед организацией вычисления, но также в процессе его развития и особенно на той стадии, когда полученные числа переводятся обратно и истолковываются на языке первоначальной задачи.*

ГЛАВА 8

ИНТЕРПОЛЯЦИЯ МНОГОЧЛЕНАМИ. ДАННЫЕ С ПРОИЗВОЛЬНЫМИ ПРОМЕЖУТКАМИ

§ 8.1. Философия

Существуют два главных применения интерполяционных формул. Прежде всего, они применяются для целей замены графически заданной функции аналитической. Изредка они используются и для того, чтобы выделить зависимость ответа от параметра, как в примере в § 1.9.

Второе главное применение их — для интерполяции в таблицах. В наши дни на универсальных цифровых вычислительных машинах интерполяцией пользуются не слишком охотно. Вместо того чтобы находить значения из таблиц, записанных в машину, предпочитают пользоваться формулами. Формулы для вычисления значений часто находят полуэмпирически и независимо от техники интерполяции многочленами. С другой стороны, высокоразвитое искусство изготовления таблиц *) делает понятным разработанность интерполяционных методов. Собирающемуся составить таблицы решительно не рекомендуется приступать к делу, не изучив прежде хорошо известную технику и возможные ловушки.

§ 8.2. Интерполяционные многочлены

Многочлен степени n

$$y(x) = a_0 + a_1x + a_2x^2 + \dots + a_nx^n = \sum_{k=0}^n a_kx^k \quad (8.2-1)$$

имеет $n+1$ коэффициент. Естественно полагать, что $n+1$ условие, наложенное на многочлен в общем виде, позволит однозначно

*, См., например, [10].

определить коэффициенты. В частности, можно потребовать, чтобы многочлен проходил через $n+1$ точку (x_i, y_i) ($i=1, 2, \dots, n+1$) с $x_i \neq x_j$. То, что многочлен проходит через точки (x_i, y_i) означает выполнение условий

$$y_i = \sum_{k=0}^n a_k x_i^k \quad (i=1, 2, \dots, n+1).$$

Определитель для этих $n+1$ линейных уравнений относительно неизвестных a_k есть определитель Вандермонда

$$W = \begin{vmatrix} 1 & x_1 & x_1^2 & \dots & x_1^n \\ 1 & x_2 & x_2^2 & \dots & x_2^n \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n+1} & x_{n+1}^2 & \dots & x_{n+1}^n \end{vmatrix} = f(x_1, x_2, \dots, x_{n+1}),$$

который, как мы сейчас покажем, не равен нулю, если $x_i \neq x_j$ для $i \neq j$.

Ясно, что определитель есть функция от x_1, x_2, \dots, x_{n+1} . Если считать его сначала функцией от x_{n+1} , то он есть многочлен степени n и обращается в нуль всякий раз, когда $x_{n+1} = x_j$ (для $j=1, 2, \dots, n$). Таким образом, $f(x_1, x_2, \dots, x_{n+1})$ содержит множители

$$\prod_{i=1}^n (x_{n+1} - x_i) = (x_{n+1} - x_1)(x_{n+1} - x_2) \dots (x_{n+1} - x_n).$$

Рассматривая определитель как функцию от x_n , мы видим точно так же, что существуют множители

$$\prod_{i=1}^{n-1} (x_n - x_i) = (x_n - x_1)(x_n - x_2) \dots (x_n - x_{n-1})$$

и вообще все множители

$$\prod_{j>i=1}^{n+1} (x_j - x_i)$$

встречаются в нашем определителе. Произведение всех этих множителей есть многочлен степени

$$n + (n-1) + (n-2) + \dots + 1 = \frac{n(n+1)}{2}.$$

Но определитель также есть многочлен той же самой степени; следовательно,

$$W = C \prod_{j>i=1}^n (x_j - x_i), \quad (8.2-2)$$

где C — некоторая константа, которую еще надо найти. Чтобы определить C , рассмотрим выражение, которое получается при умножении членов главной диагонали

$$1 \cdot x_2 \cdot x_3^2 \dots x_{n+1}^n.$$

Раскрывая произведение, мы находим точно такой же член, и, следовательно, $C = 1$. Таким образом, определитель Вандермонда не равен нулю, если $x_i \neq x_j$ для $i \neq j$.

Возвращаясь к нашей главной задаче о нахождении многочлена по $n+1$ точке (x_i, y_i) , мы видим, что ее всегда можно решить и найти коэффициенты a_k по правилу Крамера или каким-нибудь другим способом. Подставив эти значения в общий вид многочлена (8.2-1), мы можем представить результат в виде

$$\begin{vmatrix} y & 1 & x & x^2 & \dots & x^n \\ y_1 & 1 & x_1 & x_1^2 & \dots & x_1^n \\ y_2 & 1 & x_2 & x_2^2 & \dots & x_2^n \\ \dots & \dots & \dots & \dots & \dots & \dots \\ y_{n+1} & 1 & x_{n+1} & x_{n+1}^2 & \dots & x_{n+1}^n \end{vmatrix} = 0. \quad (8.2-3)$$

Впрочем, этот результат можно проверить, рассуждая следующим образом: во-первых, выражение (8.2-3) должно быть многочленом степени n по x , так как определитель можно разложить по элементам верхней строки, и, во-вторых, он проходит через $n+1$ точку (x_i, y_i) , так как подстановка этих значений в верхнюю строку делает две строки определителя одинаковыми. Возможно, что коэффициент при x^n есть нуль (условия того, чтобы это было так, достаточно ясны) и что степень многочлена меньше, чем n . Чтобы охватить и этот случай, утверждению, что многочлен имеет степень n , часто придают смысл: степень n или меньше. В вырожденном случае, когда все y_i равны, многочлен имеет степень 0 и $y(x) = C$.

Все изложенное можно подытожить, сказав, что если есть $n+1$ узловых точек функции, то можно найти многочлен степени n , который совпадает (пренебрегая ошибками округления) с функцией в узловых точках. Насколько близки обе функции между узловыми точками, будет исследовано в § 8.6, но, предположив, что они близки, мы можем использовать многочлен вместо функции в дальнейших

аналитических процессах: интегрировании, дифференцировании, отыскании нулей и т. д.

Вместо требования, чтобы многочлен проходил через некоторые данные точки, можно потребовать, чтобы он в некоторых заданных точках имел данный наклон. Так, уравнение

$$\begin{vmatrix} y & 1 & x & x^2 & x^3 \\ y_1 & 1 & x_1 & x_1^2 & x_1^3 \\ y_2 & 1 & x_2 & x_2^2 & x_2^3 \\ y'_1 & 0 & 1 & 2x_1 & 3x_1^2 \\ y'_2 & 0 & 1 & 2x_2 & 3x_2^2 \end{vmatrix} = 0$$

определяет многочлен третьей степени по x , проходящий через (x_1, y_1) с наклоном y'_1 и через (x_2, y_2) с наклоном y'_2 (так как для дифференцирования этого определителя по x достаточно продифференцировать почленно его верхнюю строку).

Если в точке дано значение y' , то вовсе не обязательно в ней должно быть указано значение y ; то же относится и к более высоким производным. Ограничением на выбор условий является требование, чтобы минор члена y обязательно не был равен нулю. В противном случае многочлена может не существовать. Это означает, что должно быть дано по крайней мере одно значение функции y_i , по крайней мере два условия на y_i и y'_i , три на y_i , y'_i , y''_i и т. д. вплоть до производной самого высокого порядка.

Только что сформулированное условие не является достаточным. Рассмотрим, например, три равноотстоящие точки, которые мы обозначим $-1, 0, 1$. В каждой точке нам даны функция y_i и вторая производная (в физических терминах положение и ускорение).

Минор при y равен

$$\begin{vmatrix} 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 2 & -6 & 12 & -20 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 2 & 6 & 12 & 20 \end{vmatrix} = 0.$$

Таким образом, эти шесть условий не определяют многочлена пятой степени.

Упражнения

8.2-1. Напишите многочлен, проходящий через (x_1, y_1) с тангенсом угла наклона y'_1 и второй производной y''_1 и имеющий тангенс угла наклона y'_2 в x_2 .

$$\text{Ответ: } \begin{vmatrix} y & 1 & x & x^2 & x^3 \\ y_1 & 1 & x_1 & x_1^2 & x_1^3 \\ y'_1 & 0 & 1 & 2x_1 & 3x_1^2 \\ y''_1 & 0 & 0 & 2 & 6x_1 \\ y'_2 & 0 & 1 & 2x_2 & 3x_2^2 \end{vmatrix} = 0.$$

8.2-2. Показать, что если использовать условия на $y, y', y'', \dots, y^{(n-1)}$ в одной точке x_1 , то многочлен будет представлять оборванный ряд Тейлора.

8.2-3. Найти нетривиальный многочлен $y(x)$ такой, что

$$y(-1) = y(0) = y(1) = y'(-1) = y'(0) = y'(1) = 0.$$

Исследовать его.

Один из возможных ответов:

$$y(x) = 3x^5 - 10x^3 + 7x.$$

(Замечание: $\sin \pi x$ также обладает этими свойствами.)

§ 8.3. Метод интерполяции Лагранжа

Другой подход к задаче интерполяции — метод Лагранжа. Основная идея этого метода состоит в том, чтобы прежде всего найти многочлен, который принимает значение 1 в одной узловой точке и 0 во всех других. Легко видеть, что функция

$$L_j(x) = \frac{(x-x_1)(x-x_2)\dots(x-x_{j-1})(x-x_{j+1})\dots(x-x_{n+1})}{(x_j-x_1)(x_j-x_2)\dots(x_j-x_{j-1})(x_j-x_{j+1})\dots(x_j-x_{n+1})} =$$

$$= \frac{\prod_{i=1}^{n+1} (x-x_i)}{\prod_{i=1}^{n+1} (x_j-x_i)}$$

(где штрих у знака произведения означает «исключая j -е значение») является требуемым многочленом степени n ; он равен 1, если $x = x_j$, и 0, когда $x = x_i$, $i \neq j$.

Многочлен $L_j(x)y_j$ принимает значение y_i в i -й узловой точке и равен нулю во всех других узлах. Из этого следует, что

$$y(x) = \sum_{i=1}^{n+1} L_i(x)y_i$$

есть многочлен степени n , проходящий через $n+1$ точку (x_i, y_i) .

Можно спросить, совпадает ли многочлен Лагранжа с многочленом (7.2-3), построенным в предыдущем параграфе. Их разность — многочлен степени не выше n и она обращается в нуль в $n+1$ узловой точке $x=x_i$, следовательно, она тождественно равна нулю. Таким образом, можно сделать следующее важное замечание: если дана $n+1$ узловая точка, то соответствующий многочлен степени n , проходящий через эти точки, однозначно (в пределах ошибок округления) определен, независимо от того, как он строится и какие обозначения использованы. Это необходимо подчеркнуть, потому что некоторые книги по численному анализу могли бы навести читателя на мысль, что различные формулы изображают разные многочлены, или из-за того, что они получены разными способами, или из-за различия в обозначениях. Если используются разные узловые точки, то, конечно, многочлены могут быть различными, но одинаковые узловые точки должны приводить к одинаковым многочленам (в пределах ошибок округления).

Интерполяционная формула Лагранжа неудобна для практического использования. Преобразования позволяют привести ее к несколько более удобному для вычисления виду. Рассмотрим частный случай, когда все $y_i = 1$. Тогда $y(x) = 1$ для всех x , так что тождественно

$$1 = \sum_{j=1}^{n+1} L_j(x).$$

Мы можем теперь разделить правую часть формулы Лагранжа на это выражение, положив

$$A_j = \frac{1}{\prod_i (x_j - x_i)};$$

после деления числителя и знаменателя на $\prod_j (x - x_j)$ получим

$$y(x) = \frac{\sum_{j=1}^{n+1} [A_j y_j (x - x_j)]}{\sum_{j=1}^{n+1} [A_j (x - x_j)]}.$$

Иногда это выражение называют «барицентрической формулой». Ее легче применять, чем формулу Лагранжа.

Рассмотрим теперь задачу нахождения интерполяционного многочлена, который в каждой точке x_i принимает значение y_i и имеет производные $y_i^{(l)}$ ($l=1, \dots, n+1$). Такой многочлен называется многочленом Эрмита. Ясно, что он имеет степень $2n+1$ (так как должны быть удовлетворены $2n+2$ условия).

Нас больше интересует метод, применяемый при построении многочлена, чем окончательный результат. Этот метод есть естественное продолжение метода Лагранжа. Мы ищем такие многочлены $H_i(x)$ и $h_i(x)$ степени не выше $2n+1$, что при $i \neq j$

$$\begin{aligned} H_j(x_i) &= 0, & h_j(x_i) &= 0, \\ H_j'(x_i) &= 0, & h_j'(x_i) &= 0, \\ H_j(x_j) &= 1, & h_j(x_j) &= 0, \\ H_j'(x_j) &= 0, & h_j'(x_j) &= 1. \end{aligned}$$

После того как эти многочлены будут найдены, решение поставленной задачи можно будет записать в виде

$$y(x) = \sum_{j=1}^{n+1} [H_j(x) y_j + h_j(x) y_j'].$$

Прежде всего построим $h_j(x)$. Если $h_j(x_i) = 0$ и $h_j'(x_i) = 0$, то $h_j(x)$ должен содержать множитель

$$(x - x_i)^2 \quad (\text{для } i \neq j).$$

Если $h_j(x_j) = 0$, но $h_j'(x_j) \neq 0$, то существует простой множитель $(x - x_j)$. Чтобы получить $h_j(x_j) = 1$, следует взять

$$h_j(x) = \frac{(x - x_1)^2 (x - x_2)^2 \dots (x - x_{j-1})^2 (x - x_j) (x - x_{j+1})^2 \dots (x - x_{n+1})^2}{(x_j - x_1)^2 (x_j - x_2)^2 \dots (x_j - x_{j-1})^2 (x_j - x_{j+1})^2 \dots (x_j - x_{n+1})^2}.$$

Читатель легко убедится, что $h_j(x)$ обладает требуемыми свойствами.

Построение $H_j(x)$ почти такое же. Опять мы имеем множители

$$(x - x_i)^2 \quad (\text{для } i \neq j),$$

но в то же время требуем, чтобы

$$H_j(x_j) = 1, \quad H_j'(x_j) = 0.$$

Будем искать $H_j(x)$ в виде

$$H_j(x) = \frac{(x - x_1)^2 (x - x_2)^2 \dots (x - x_{j-1})^2 (ax + b) (x - x_{j+1})^2 \dots (x - x_{n+1})^2}{(x_j - x_1)^2 (x_j - x_2)^2 \dots (x_j - x_{j-1})^2 (x_j - x_{j+1})^2 \dots (x_j - x_{n+1})^2}.$$

Условие $H_j(x_j) = 1$ требует, чтобы

$$ax_j + b = 1.$$

Взяв производную от H_j и положив в ней $x = x_j$, получим

$$\begin{aligned} H_j'(x_j) &= \frac{2}{x_j - x_1} + \frac{2}{x_j - x_2} + \dots + \frac{2}{x_j - x_{j-1}} + \\ &\quad + a + \frac{2}{x_j - x_{j+1}} + \dots + \frac{2}{x_j - x_{n+1}} = 0. \end{aligned}$$

Таким образом,

$$a = -2 \sum_{i=1}^{n+1} \frac{1}{x_j - x_i}, \quad b = 1 - ax_j$$

и требуемая формула построена.

Этот метод построения многочленов, обладающих специфическими свойствами, используется в теории численных методов весьма часто и им следует овладеть. Мы применили его в двух конкретных случаях — при построении интерполяционных полиномов Лагранжа и Эрмита; но читатель должен уяснить себе общий метод подбора $y_i, y'_i, \dots, y_i^{(n)}$ в x_i . Как будет видно при исследовании погрешности интерполяционных формул, использование производных в близких точках весьма увеличивает точность формулы.

Упражнение 8.3-1. Указать способ получения интерполяционной формулы, которая имеет значения y_i, y'_i и y''_i в каждой точке x_i .

§ 8.4. Интерполяционная формула Ньютона

Рассмотренные методы нахождения интерполяционного многочлена по $n+1$ точке подразумевали, что множество используемых узлов известно. Часто известным является лишь требуемая точность, а множество узлов, которые следует использовать, определяется информацией о том, как вычисляется функция. Интерполяционная формула Ньютона, которая будет изложена, представляет собой просто другой способ написания интерполяционного многочлена. Она полезна, потому что число используемых узлов может быть легко увеличено или уменьшено без повторения всего вычисления. Формула Ньютона из § 1.5 есть просто ее частный случай, когда узловые точки равноудалены.

Как и раньше, обозначим многочлен, проходящий через $n+1$ точку (x_i, y_i) ($i = 1, 2, \dots, n+1$), через $P_n = P_n(x)$. Можно написать

$$P_n(x) = y_1 + (x - x_1) P_{n-1}(x),$$

где $P_{n-1}(x)$ — некоторый многочлен степени $n-1$. Ясно, что $P_n(x_1) = y_1$ и следует заботиться лишь об n точках ($i = 2, 3, \dots, n+1$). Предыдущее уравнение может быть записано в виде

$$P_{n-1}(x) = \frac{P_n(x) - y_1}{x - x_1},$$

и, следовательно, требуется, чтобы

$$P_{n-1}(x_i) = \frac{P_n(x_i) - P_n(x_1)}{x_i - x_1} = \frac{y_i - y_1}{x_i - x_1} \quad (i = 2, 3, \dots, n+1).$$

Тем самым нужно, чтобы $P_{n-1}(x)$ проходил через точки

$$\left(x_i, \frac{y_i - y_1}{x_i - x_1}\right) \quad (i = 2, 3, \dots, n+1).$$

Величины

$$\frac{y_i - y_1}{x_i - x_1} = [x_i, x_1] = [x_1, x_i]$$

называются *разделенными разностями* и обычно записываются в квадратных скобках *).

На следующем шаге надо написать

$$P_{n-1}(x) = [x_1, x_2] + (x - x_2) P_{n-2}(x)$$

и требовать, чтобы $P_{n-2}(x)$ принимал значения разделенных разностей от разделенных разностей

$$[[x_i, x_1], [x_2, x_1]] = \frac{[x_i, x_1] - [x_2, x_1]}{x_i - x_2} = [x_i, x_2, x_1].$$

Нетрудно видеть, что разделенные разности первого порядка не зависят от порядка аргументов в квадратных скобках. Покажем теперь, что это верно и для разделенных разностей второго порядка. Если начать с трех точек

$$(x_1, y_1), (x_2, y_2), (x_3, y_3),$$

то получим единственный многочлен второй степени, проходящий через три точки. Его можно записать так:

$$y = y_1 + (x - x_1) \{ [x_2, x_1] + (x - x_2) [x_3, x_2, x_1] \}.$$

Если теперь взять точки в таком порядке: $(x_a, y_a), (x_b, y_b), (x_c, y_c)$, то получим

$$y = y_a + (x - x_a) \{ [x_b, x_a] + (x - x_b) [x_c, x_b, x_a] \}.$$

Так как оба этих уравнения определяют один и тот же квадратный трехчлен, то коэффициенты при x^2 должны быть одинаковы, так что два символа

$$[x_3, x_2, x_1] = [x_c, x_b, x_a]$$

выражают одно и то же.

Вообще определяем

$$[x_1, x_2, x_3, \dots, x_n] = \frac{[x_1, x_2, \dots, x_{n-2}, x_{n-1}] - [x_1, x_2, \dots, x_{n-2}, x_n]}{x_{n-1} - x_n}$$

и в точности таким же способом показываем, что разности не зависят от порядка x_i . Заметим, что знаменатель есть разность неповторяющихся значений x , взятых в том же порядке.

*) Используются и другие обозначения, например $f[x_i, x_1]$ и $\rho[x_i, x_1]$.

Один из способов построения таблицы необходимых значений следующий:

x_1	y_1^*			
		$[x_2, x_1]^*$		
x_2	y_2		$[x_3, x_2, x_1]^*$	
		$[x_3, x_1]$		$[x_4, x_3, x_2, x_1]^*$
x_3	y_3		$[x_4, x_3, x_1]$	
		$[x_4, x_1]$
x_4	y_4	...		
...	...			

Звездочки означают числа, использующиеся как опорные значения при вычислении значений многочлена.

Из этой таблицы можно написать

$$y(x) = y_1 + (x - x_1)[x_2, x_1] + (x - x_2)\{[x_3, x_2, x_1] + (x - x_3)[\dots]\}.$$

В качестве примера рассмотрим таблицу логарифмов

x	$\lg x$	[,]	[,,]	[,,,]
1	0,0000*			
2	0,3010	0,30100*		
3	0,4771	0,23855	-0,06245*	
4	0,6021	0,20070	-0,05015	+0,01230*

Звездочки означают опорные значения. Отсюда получаем

$$y(x) = 0 + (x - 1)\{0,3010 + (x - 2)[(-0,06245) + (x - 3)(0,01230)]\}.$$

В частности,

$$y(2,5) = \frac{3}{2} \left\{ 0,3010 + \frac{1}{2} \left[(-0,06245) + \left(-\frac{1}{2}\right)(0,01230) \right] \right\} = 0,40001.$$

Верное значение

$$\lg 2,5 = 0,3979.$$

Упражнение 8.4-1. Укажите способ добавления еще одной данной точки в конце таблицы разделенных разностей. Покажите также, как добавить одну точку вверх. В первом случае какое изменение следует сделать в многочлене Ньютона?

§ 8.5. Другая форма для таблицы разделенных разностей

Таблица разделенных разностей, которая лежит в основе интерполяционной формулы Ньютона, может быть написана в другой, иногда более полезной форме. Заметим, что

$$[x_1, x_2, \dots, x_n] = \frac{[x_1, x_2, \dots, x_{n-1}] - [x_1, x_2, \dots, x_{n-2}, x_n]}{x_{n-1} - x_n} = \frac{[x_1, x_2, \dots, x_{n-1}] - [x_2, x_3, \dots, x_n]}{x_1 - x_n}.$$

В частности,

$$[x_1, x_2, x_3] = \frac{[x_1, x_2] - [(x_1, x_3)]}{x_2 - x_3} = \frac{[x_1, x_2] - [x_3, x_3]}{x_1 - x_3}.$$

Таким образом, можно написать

	[,]	[,]	[,,]
x_1	y_1		
	$[x_2, x_1]$		
x_2	y_2	$[x_3, x_2, x_1]$	
	$[x_3, x_2]$		$[x_4, x_3, x_2, x_1]$
x_3	y_3	$[x_4, x_3, x_2]$	
x_4	y_4	$[x_4, x_3]$	

Несмотря на то, что числа в этой таблице отличаются от чисел в предыдущем разделе, числа в верхнем ряду, которые только и используются в формуле Ньютона, те же самые.

Взяв тот же самый пример для $\lg x$, получаем

x	$\lg x$	[,]	[,,]	[,,,]
1	0,0000			
2	0,3010	0,3010		
3	0,4771	0,1761	-0,06245	
4	0,6021	0,1250	-0,02555	+0,01230

В таком виде мы легко можем добавить строку и к тому и к другому концу таблицы и надеяться, что числа в таблице изменятся гладко. Хотя теоретически нет необходимости располагать точки в порядке возрастания (или убывания) x_j , гладкость в таблице нарушается, если такого порядка нет. Первые разделенные разности — се-

кущие линии и, следовательно, близки к первым производным в интервале (x_n, x_{n+1}) . Аналогично вторые разделенные разности являются локальными приближениями ко вторым производным и т. д.

Упражнения

8.5-1. Используя таблицу

x	2	0	3	1
y	8	0	27	1

вычислить интерполяционный многочлен Ньютона. Показать, что $y = x^3$.

8.5-2. Вычислить таблицу разностей обоими способами (§§ 8.4 и 8.5) для $y = \sin x$ с шагом в 30° :

x	0	30°	60°	90°
$\sin x$	0,0000	0,5000	0,8660	1,0000

Найти аппроксимационный многочлен.

8.5-3. Как прибавить одну точку внизу таблицы и вычислить следующую разделенную разность, если в памяти вычислительной машины хранится лишь верхний ряд таблицы разделенных разностей? Как прибавить одну точку наверху и найти новую строку?

§ 8.6. Погрешность многочленной аппроксимации

Задав функцию $f(x)$, мы брали $n+1$ узловую точку (x_i, y_i) ($i=1, 2, \dots, n+1$) и находили многочлен $P_n(x)$, проходящий через эти точки. Затем мы собирались использовать многочлен вместо первоначальной функции, и поэтому важно рассмотреть вопрос о том, как сильно могут отличаться функция и многочлен в точках, отличных от узловых (где они совпадают в пределах ошибки округления).

В качестве примера рассмотрим функцию $y(x) = \lg x$. В §§ 8.4 и 8.5 мы находили многочлен, аппроксимирующий функцию. Исследуем значения разности $\lg x - P(x)$ в узловых точках и средних точках между ними, где можно ожидать, что она довольно велика.

x	$\lg x$	$P(x)$	$\lg x - P(x)$	x	$\lg x$	$P(x)$	$\lg x - P(x)$
1,0	0,0000	0,0000	0,0000	3,0	0,4771	0,4771	0,0000
1,5	0,1761	0,1707	0,0054	3,5	0,5441	0,5414	0,0027
2,0	0,3010	0,3010	0,0000	4,0	0,6021	0,6021	0,0000
2,5	0,3979	0,4000	-0,0021				

Теоретическое выражение для разности между первоначальной функцией $f(x)$ и аппроксимационным многочленом $P(x)$ может быть найдено, если заметить, что разность равна нулю во всех узловых точках, и написать

$$y(x) - P(x) = (x - x_1)(x - x_2) \dots (x - x_{n+1}) K(x),$$

где $K(x)$ выбрано соответствующим образом. Для произвольного x^* имеем

$$y(x^*) - P(x^*) = (x^* - x_1)(x^* - x_2) \dots (x^* - x_{n+1}) K(x^*) = 0.$$

Теперь рассмотрим функцию

$$\Phi(x) = y(x) - P(x) - (x - x_1)(x - x_2) \dots (x - x_{n+1}) K(x^*).$$

Если $y(x)$ имеет $(n+1)$ -ю производную, то функцию $\Phi(x)$ можно продифференцировать $n+1$ раз. Так как $P(x)$ — многочлен степени n , а $K(x^*)$ — константа, то

$$\Phi^{(n+1)}(x) = y^{(n+1)}(x) - (n+1)! K(x^*).$$

Но $\Phi(x)$ обращается в нуль $n+2$ раза (в точках x^* и x_1, x_2, \dots, x_{n+1}). Следовательно, по теореме о среднем значении $\Phi'(x)$ обращается в нуль по крайней мере $n+1$ раз в интервале, содержащем все значения (включая x^*). Продолжая применять эту теорему, находим, что $\Phi^{(n+1)}(x)$ обращается в нуль по крайней мере $n+2-k$ раз и что $\Phi^{(n+1)}(x)$ обращается в нуль по крайней мере однажды. Таким образом, в интервале значений x существует такое \bar{x} , что

$$y^{(n+1)}(\bar{x}) = (n+1)! K(x^*).$$

Отсюда можно получить значение константы $K(x^*)$. Подставив его в первоначальное выражение, получаем

$$y(x^*) = P(x^*) + \frac{(x^* - x_1)(x^* - x_2) \dots (x^* - x_{n+1}) y^{(n+1)}(\bar{x})}{(n+1)!}.$$

Так как x^* произвольно, то вместо x^* можно написать x :

$$y(x) = P(x) + \frac{(x - x_1)(x - x_2) \dots (x - x_{n+1}) y^{(n+1)}(\bar{x})}{(n+1)!}. \quad (8.6-1)$$

Используя это выражение, чтобы оценить ошибку, сделанную при интерполяции в таблице логарифмов, получим

$$\frac{(x-1)(x-2)(x-3)(x-4)}{4!} y^{(IV)}(\bar{x}).$$

Оценивая ошибку при $x = \frac{3}{2}$, получаем

$$\frac{(\frac{1}{2})(-\frac{1}{2})(-\frac{3}{2})(-\frac{5}{2})}{4!} \cdot \frac{6}{x^4} = \frac{15}{64} \cdot \frac{1}{x^4}.$$

О значении \bar{x} известно лишь, что $1 \leq \bar{x} \leq 4$. В худшем случае ошибка равна $\frac{15}{64} \approx 0,23$, тогда как в лучшем случае 0,001; в действительности же она около 0,0054, что показывает, с какой маленькой точностью мы можем оценить ошибку, используя полученное выражение.

Следует заметить, что среднее значение $\bar{x} = \bar{x}(x)$ зависит от x и в действительности для какого-либо значения x может существовать несколько значений \bar{x} . Следовательно, \bar{x} не обязательно непрерывная функция от x . Этот последний факт иллюстрируется применением теоремы о среднем значении к функции $y(x) = x(1-x)^2$ и выбором $a=0$ в обычной форме теоремы о среднем значении:

$$y'(\bar{x}) = \frac{y(x) - y(a)}{x - a} = \frac{y(x)}{x} = (1-x)^2.$$

Таким образом, $y'(\bar{x}) \geq 0$ и при x , меняющемся от 0 до 1, \bar{x} меняется от 0 до $\frac{1}{3}$, при x , меняющемся от 1 до 2, \bar{x} меняется от $\frac{1}{3}$ до 0, а когда x становится больше 2, перепрыгивает через интервал, где $y' < 0$, к $\bar{x} > 1$. Следовательно, $\bar{x} = \bar{x}(x)$ не непрерывная функция от x .

Исследуем теперь частный случай интерполирования по формуле Эрмита, где и функция и производная заданы в каждой из $n+1$ узловых точек. Чтобы найти ошибку, заметим, что $y(x) - P(x)$ имеет двойной нуль в каждой x_i и можно положить

$$y(x) - P(x) = [(x - x_1)(x - x_2) \dots (x - x_{n+1})]^2 K(x).$$

Продолжая как и выше, находим

$$y(x) = P(x) + \frac{(x - x_1)^2 (x - x_2)^2 \dots (x - x_{n+1})^2}{(2n+2)!} y^{(2n+2)}(\bar{x}). \quad (8.6-2)$$

Детали выкладок предоставляем сделать читателю.

Форма остаточного члена, которой мы пользовались, соответствует остаточному члену в ряде Тейлора, известному как остаточный член в форме Лагранжа.

Упражнения

8.6-1. Найти остаточный член в случае, когда в каждой узловой точке известны функция и первые две производные (см. упражнение 8.3-1).

8.6-2. Показать, что остаточный член в случае, когда в x_i определены функция и первые $m_i - 1$ производные, равен

$$(x - x_1)^{m_1} (x - x_2)^{m_2} \dots (x - x_{n+1})^{m_{n+1}} \frac{y^{(m)}(\bar{x})}{m!}, \quad (8.6-3)$$

где $m = m_1 + m_2 + \dots + m_{n+1}$.

(З а м е ч а н и е: это не самый общий случай; в любой точке, где дана производная, мы также требуем, чтобы были заданы и все производные более низкого порядка. В § 8.2 мы так не делали; насколько известно автору, формулы остаточного члена в общем случае опубликовано не было.)

§ 8.7. Трудности приближения многочленом

Принято, как это сделано в предыдущем параграфе, выражать ошибку приближения многочленом в терминах соответствующей производной от функции, которая была аппроксимирована. Обычно думают, что «для большинства разумных функций» такие выражения для ошибки становятся маленькими для достаточно большого n . К сожалению, это не так.

Ограничим наше рассуждение аналитическими функциями, т. е. функциями, имеющими сходящийся ряд Тейлора

$$y(x) = \sum_{n=0}^{\infty} \frac{(x-x_0)^n}{n!} y^{(n)}(x_0)$$

во всех точках x_0 интересующей нас области. Если к тому же функция целая, т. е. ее ряд сходится всюду в конечной части комплексной плоскости, как для $\sin x$, $\exp x$, многочлена от x и т. д., то, действительно, все производные достаточно высокого порядка малы. Но если функция имеет особенность в конечной части комплексной плоскости, как $\operatorname{tg} x$, $\log x$, рациональная функция от x и т. д., то ряд Тейлора должен иметь конечный радиус сходимости R , а это в свою очередь означает, что для бесконечного числа значений n

$$\frac{(R+\varepsilon)^n}{n!} |y^{(n)}(x_0)| \geq 1 \quad (\varepsilon > 0)$$

и

$$|y^{(n)}(x_0)| \geq \frac{n!}{(R+\varepsilon)^n}.$$

Другими словами, верхняя грань n -й производной растет как $n!$. В качестве примера рассмотрим

$$\begin{aligned} y &= \ln x, \\ y' &= \frac{1}{x}, \\ y'' &= -\frac{1}{x^2}, \\ y''' &= \frac{2!}{x^3}, \\ &\dots \dots \dots \\ y^{(n)} &= \frac{(-1)^{n-1} (n-1)!}{x^n}. \end{aligned}$$

Таким образом, даже если кривая $y = \ln x$ выглядит гладкой вблизи некоторых значений x , тем не менее, когда n становится

большим, производные в этой точке делаются очень большими по величине и ведут себя как $n!$.

Это — общий случай: для «большинства функций» некоторые из производных более высокого порядка имеют тенденцию расти как $n!$. Ограниченными производными обладают лишь некоторые целые функции *). Даже у производных от многочленов есть тенденция расти по величине до n -й производной, которая равна $a_n n!$; после нее все производные становятся нулями. Конечно, у функции многие из производных высокого порядка могут быть маленькими. Например, четная функция имеет все свои производные нечетного порядка равными нулю, но если функция не является целой, то существует также бесконечное число производных четного порядка, которые имеют тенденцию вести себя как $n!$.

Было бы приятно, если бы приходилось иметь дело лишь с целыми функциями, которые обладают ограниченными производными, но в действительности оказывается, что если функция является целой, то весьма вероятно, что вся проблема может быть решена аналитически, тогда как если необходимо применять численные методы, то функция, как правило, ведет себя несколько хуже.

В качестве обоснования многочленных приближений часто цитируется ([20], стр. 19) теорема Вейерштрасса, которая утверждает, грубо говоря, что непрерывная функция на замкнутом отрезке может быть равномерно приближена многочленами. Однако метод точного совпадения в узловых точках, которым мы воспользовались, не является способом, которым определяются многочлены Вейерштрасса; следовательно, теорема, хотя, возможно, и многообещающая, не применима.

Хорошо известным примером является простая функция ([39], стр. 35—39) $y(x) = \frac{1}{1+x^2}$. С ростом числа равноотстоящих узлов отклонение аппроксимационного многочлена от функции между некоторыми узлами не уменьшается. Таким образом, даже для равноотстоящих узлов нельзя полагаться на то, что многочлен будет хорошей аппроксимацией, если единственным требованием, которому должен удовлетворять такой многочлен, является точное совпадение в узловых точках. Объясняется этот факт, конечно, тем, что производные растут слишком быстро.

В качестве примера рассмотрим функцию Римана, заданную таблицей 8.7-1. Верхний ряд разностей не имеет тенденции быстро падать, что происходит из-за очевидной особенности при $x=1$.

Следует заметить однако, что если таблица дана через полшага по отношению к другой таблице той же самой функции, то ее пер-

*) Обратное неверно: производные целой функции не обязательно ограничены. Например, если $y = xe^x$, то $y^{(n)}(0) = n + 1$.

вые разности вдвое меньше, вторые в четыре раза меньше и т. д., чем разности второй таблицы. Это наводит на мысль об эмпирическом правиле: если разности в таблице стремятся к нулю быстро, то, вероятно, был взят слишком маленький шаг, а таблица слишком велика, тогда как, если разности не становятся маленькими, следует рассмотреть более мелкий шаг.

Таблица 8.7-1

Дзета-функция Римана $\zeta(x) = \sum_{k=1}^{\infty} \frac{1}{k^x}$

x	$\zeta(x)$	$\Delta\zeta$	$\Delta^2\zeta$	$\Delta^3\zeta$	$\Delta^4\zeta$	$\Delta^5\zeta$
2	1,64493					
3	1,20206	—0,44287	0,32313			
4	1,08232	—0,11974	7435	—0,24878	0,20023	
5	1,03693	—0,04539	2580	—4855	3335	—0,16688
6	1,01734	—0,01959	1060	—1520	932	—2403
7	1,00835	—0,00899	472	—588	336	—596
8	1,00408	—0,00427	220	—252	137	—199
9	1,00201	—0,00207	105	—115	62	—75
10	1,00099	—0,00102	52	—53	27	—35
11	1,00049	—0,00050	26	—26	11	—16
12	1,00025	—0,00024	11	—15	11	—0
13	1,00012	—0,00013	7	—4	0	—11
14	1,00006	—0,00006	3	—4	3	+3
15	1,00003	—0,00003	2	—1	—1	—4
16	1,00002	—0,00001	0	—2		
17	1,00001	—0,00001				

Последнее и, возможно, самое серьезное возражение против аппроксимации многочленами заключается в том, что она редко имеет какое-либо физическое значение, которое приводит к полезным представлениям.

С другой стороны, теория аппроксимации многочленами проста, хорошо развита, требует минимума вычислений и полезна. Опыт показывает, что приближение многочленами во многих случаях дает хороший результат, хотя остаточный член часто либо вообще трудно оценить, либо его оценка очень пессимистична. Значения таблицы разностей могут быть весьма похожими на значения производной, и они показывают, насколько велик вклад от увеличения числа членов в интерполяционном многочлене Ньютона. Однако использование разностей как индикатора значений производной иногда опасно; так, для целого числа x все значения $\sin \pi x$ равны нулю, так что таблица разностей, также состоящая из нулей, заставляет предположить нулевую ошибку в аппроксимации $\sin \pi x = 0$, что едва ли верно. Здесь ошибка очевидна, но в некоторых случаях она может быть пропущена.

Упражнение 8.7-1. Вычислить n -ю производную от

$$\frac{1}{1+x^2} = \frac{1}{2i} \left(\frac{1}{x-i} - \frac{1}{x+i} \right) \quad (i = \sqrt{-1}).$$

§ 8.8. О выборе узловых точек

Проблема того, какие узловые точки выбирать, возникает всякий раз, когда приходится интерполировать. Если бы мы имели полную свободу выбора, то выбрали бы именно то значение x , для которого собираемся интерполировать, сделав, таким образом, задачу тривиальной. Но часто случается, что мы имеем обширную группу значений (x_i, y_i) , которые можем использовать и ни одно из которых не совпадает с желаемым значением x . Если о величине производной, встречающейся в выражении для ошибки, ничего не известно, то можно лишь выбрать наши узловые точки так, чтобы минимизировать коэффициент (см. уравнение (8.6-3) в упражнении 8.6-2)

$$(x - x_1)^{m_1} (x - x_2)^{m_2} \dots (x - x_n)^{m_n}$$

перед выражением для производной.

Здравый смысл и минимизация этого множителя требуют одного и того же: использования информации, близкой к месту, в котором нужно интерполировать. При применении интерполяционной формулы Ньютона

$$y(x) = y(x_1) + (x - x_1) \{ [x_2, x_1] + (x - x_2) \{ [x_3, x_2, x_1] \} \} \dots$$

возникает желание держать x близким именно к x_1 . Это желание следует сдерживать; выгоднее равновесие всех множителей, входящих в ошибку. Таким образом, следует выбирать значения x_i сгруппированными, если возможно, по обе стороны от требуемого значения x .

Иногда возникает другой вопрос (ответ на который не может быть дан здесь просто). Допустим, что мы хотим интерполировать значения на всем интервале. Где нужно выбрать узловые точки, чтобы минимизировать максимальную ошибку? Ответ на этот вопрос дается теорией многочленов Чебышева, изложенной в гл. 19.

ГЛАВА 9

ИНТЕРПОЛЯЦИЯ МНОГОЧЛЕНАМИ. РАВНООТСТОЯЩИЕ УЗЛЫ

§ 9.1. Формула Ньютона для интерполирования

Очень часто имеющаяся информация о функции (в виде значений в узловых точках) задана на множестве равноотстоящих значений x . В этом случае большая часть формул, вычислений, как, впрочем, и затрагиваемых идей, заметно упрощается.

Для разностей при равноотстоящих значениях обычно применяют обозначения первой части:

$$\Delta y_n = y_{n+1} - y_n.$$

Это — обычное для исчисления разностей обозначение, но $\Delta x = h$ здесь фиксировано:

$$\Delta x_n = x_{n+1} - x_n = h.$$

Мы имеем также

$$\Delta^2 y_n = y_{n+2} - 2y_{n+1} + y_n$$

и т. д.

Эти разности соответствуют разделенным разностям

$$[x_2, x_1] = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\Delta y_1}{h},$$

$$[x_3, x_2, x_1] = \frac{[x_3, x_2] - [x_2, x_1]}{x_3 - x_1} = \frac{\frac{\Delta y_2}{h} - \frac{\Delta y_1}{h}}{2h} = \frac{\Delta^2 y_1}{2!h^2}$$

и вообще

$$[x_1, x_2, \dots, x_n] = \frac{\Delta^{n-1} y_1}{(n-1)! h^{n-1}}.$$

Разности приближают производные в точках, лежащих посередине между используемыми узлами:

$$\begin{aligned} \Delta y_1 &\sim h \frac{dy\left(x + \frac{h}{2}\right)}{dx}, \\ \Delta^2 y_1 &\sim h^2 \frac{d^2 y\left(x + \frac{h}{2}\right)}{dx^2}, \\ &\dots \dots \dots \\ \Delta^n y_1 &\sim h^n \frac{d^n y\left(x + \frac{nh}{2}\right)}{dx^n}. \end{aligned}$$

Эти соотношения дают возможность аппроксимировать выражения в дифференциальном исчислении конечноразностными.

Формула Ньютона в этих новых обозначениях для равноотстоящих узловых точек имеет вид (см. уравнение (1.5-4) и § 8.4)

$$y = y_0 + (x - x_0) \frac{\Delta y_0}{n} + (x - x_0)(x - x_0 - h) \frac{\Delta^2 y_0}{2h^2} + \\ + (x - x_0)(x - x_0 - h)(x - x_0 - 2h) \frac{\Delta^3 y_0}{3! h^3} + \dots$$

Полагая $x_0 = 0$, получим

$$y = y_0 + x \frac{\Delta y_0}{h} + x(x - h) \frac{\Delta^2 y_0}{2h^2} + x(x - h)(x - 2h) \frac{\Delta^3 y_0}{3! h^3} + \dots,$$

и если дальше положить $h = 1$, то получается (ср. с уравнением (1.5-4)):

$$y = y_0 + x \Delta y_0 + x(x - 1) \frac{\Delta^2 y_0}{2!} + x(x - 1)(x - 2) \frac{\Delta^3 y_0}{3!} + \dots$$

Упражнение 9.1-1. Напишите остаточный член формулы Ньютона, где информация задана в точках $-n, -(n-1), \dots, 0, 1, 2, \dots, n$ и используются разности $2n$ -го порядка. Оцените максимум коэффициента при производной.

§ 9.2. Интерполирование в таблицах

Одним из главных применений интерполяционных формул с равным шагом является их применение для интерполирования в таблицах с равноотстоящими аргументами. Типичным примером является интеграл ошибок, таблица разностей которого дана в таблице 9.2-1. Разности ведут себя хорошо в том смысле, что они стремятся к нулю, оставаясь довольно гладкими в столбце третьих разностей. Начиная с четвертых разностей и особенно в пятых разностях преобладает уже шум округления. Поэтому для этой таблицы нельзя надеяться, применяя какую-нибудь интерполяционную формулу, идти далее четвертых разностей.

В принципе, интерполяционный многочлен может быть получен по формуле Ньютона для интерполирования и, не считая эффектов округления, ответ будет точно таким же, какой получился при использовании интерполяционного многочлена, полученного другим способом, использующим те же самые узловые точки. Тем не менее на практике применяется много других формул, и для общего образования, чтобы читатель мог понять методы, приводящиеся в предисловиях к большинству таблиц, мы разберем некоторые из них, наиболее популярные.

Таблица 9.2-1

Интеграл ошибок $\frac{1}{\sqrt{2\pi}} \int_0^z e^{-\frac{\theta^2}{2}} d\theta$

t	$f(t)$	Δ	Δ^2	Δ^3	Δ^4	Δ^5
0,00	0,0000					
0,25	0,0987	987				
		928	— 59	—50		
0,50	0,1915	819	—109	—31	+19	
0,75	0,2734	679	—140	— 8	+23	+ 4
1,00	0,3413	531	—148	+ 5	+13	—10
1,25	0,3944	388	—143	+22	+17	+ 4
1,50	0,4332	267	—121	+28	+ 6	—11
1,75	0,4599	174	— 93	+24	— 4	—10
2,00	0,4773	105	— 69	+24	0	+ 4
2,25	0,4878	60	— 45	+17	— 7	— 7
2,50	0,4938	32	— 28	+13	— 4	+ 3
2,75	0,4970	17	— 15	+ 5	— 8	— 4
3,00	0,4987	7	— 10	+ 7	+ 2	+10
3,25	0,4994	4	— 3	0	— 7	— 9
3,50	0,4998	1	— 3	+ 2	+ 2	+ 9
3,75	0,4999	1	— 1			
4,00	0,5000					

§ 9.3. Ромбовидная диаграмма

Ромбовидная диаграмма является приспособлением для доказательства того, что огромное число формул, которые кажутся различными, в действительности все одинаковы.

Мы определили биномиальные коэффициенты в § 1.3:

$$C(u+k, n) = \frac{(u+k)(u+k-1)(u+k-2) \dots (u+k-n+1)}{n!}.$$

Числитель и знаменатель этой формулы содержат по n сомножителей. Коэффициент $C(u+k, n)$, если его рассматривать как функцию от u , есть многочлен степени n . На рис. 9.3-1 изображена ромбовидная диаграмма. Линия, начинаясь в точке на левом краю и следуя по

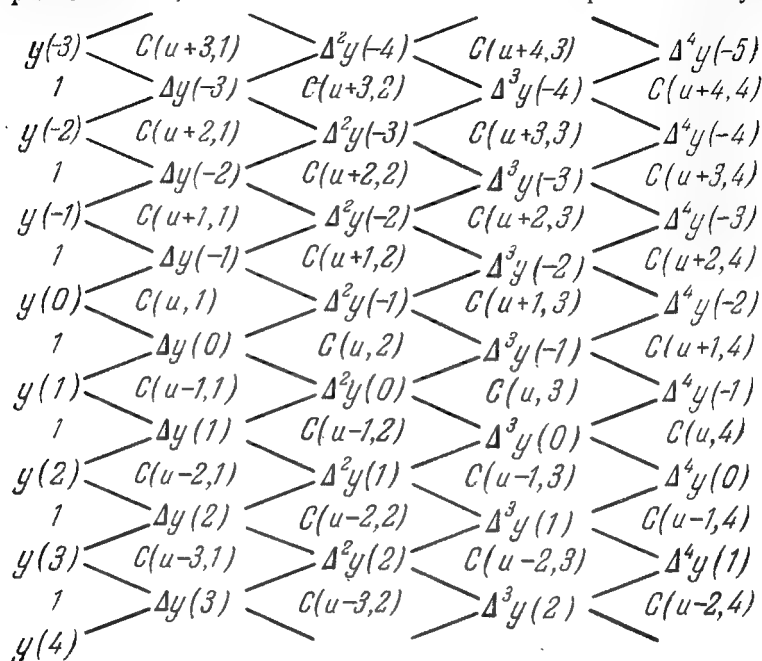


Рис. 9.3-1. Ромбовидная диаграмма.

какой-либо дорожке поперек страницы, определяет интерполяционную формулу, если применять следующие правила:

- 1а. Для шага слева направо — складывать.
- 1б. Для шага справа налево — вычитать.
- 2а. Если тангенс угла наклона шага положительный, то использовать произведение разделенной разности на коэффициент, находящийся под ней.
- 2б. Если тангенс угла наклона шага отрицательный, то использовать произведение разделенной разности на коэффициент, находящийся над ней.
- 3а. Если шаг горизонтальный и проходит через разность, брать произведение разности на среднее арифметическое из коэффициентов, расположенных выше и ниже ее.
- 3б. Если шаг горизонтальный и проходит через коэффициент, брать произведение коэффициента на среднее арифметическое из разностей, расположенных выше и ниже его.

В качестве примера применения правил 1а и 2б рассмотрим путь, начинающийся с $y(0)$ и идущий направо вниз. Мы получаем формулу

$$y(u) = y(0) + C(u, 1) \Delta y(0) + C(u, 2) \Delta^2 y(0) + C(u, 3) \Delta^3 y(0) + \dots = \\ = y(0) + u \Delta y(0) + \frac{u(u-1)}{2} \Delta^2 y(0) + \frac{u(u-1)(u-2)}{3!} \Delta^3 y(0) + \dots,$$

которая является формулой Ньютона. Если бы мы пошли направо вверх, мы применили бы правила 1а и 2а и получили бы вторую формулу Ньютона

$$y(u) = y(0) + C(u, 1) \Delta y(-1) + C(u+1, 2) \Delta^2 y(-2) + \\ + C(u+2, 3) \Delta^3 y(-3) + \dots = y(0) + u \Delta y(-1) + \\ + \frac{(u+1)u}{2} \Delta^2 y(-2) + \frac{(u+2)(u+1)u}{3!} \Delta^3 y(-3) + \dots \quad (9.3-1)$$

Чтобы получить формулу Стирлинга, начнем с $y(0)$ и пойдем горизонтально направо, применяя правила 3а и 3б:

$$y(u) = y(0) + u \frac{\Delta y_0 + \Delta y_{-1}}{2} + \frac{C(u+1, 2) + C(u, 2)}{2} \Delta^2 y_{-1} + \\ + C(u+1, 3) \frac{\Delta^3 y_{-2} + \Delta^3 y_{-1}}{2} + \dots = y_0 + u \frac{\Delta y_0 + \Delta y_{-1}}{2} + \\ + \frac{u^2}{2} \Delta^2 y_{-1} + \frac{u(u^2-1)}{3!} \frac{\Delta^3 y_{-2} + \Delta^3 y_{-1}}{2} + \dots \quad (9.3-2)$$

Если мы начнем посередине между $y(0)$ и $y(1)$, то получим формулу Бесселя

$$y(u) = 1 \frac{y_0 + y_1}{2} + \frac{C(u, 1) + C(u-1, 1)}{2} \Delta y_0 + C(u, 2) \frac{\Delta^2 y_{-1} + \Delta^2 y_0}{2} + \dots \\ \dots = \frac{y_0 + y_1}{2} + \left(u - \frac{1}{2}\right) \Delta y_0 + \frac{u(u-1)}{2} \cdot \frac{\Delta^2 y_{-1} + \Delta^2 y_0}{2} + \dots \quad (9.3-3)$$

Если мы пойдем зигзагом, соответствующим образом, то можно получить интерполяционную формулу Гаусса

$$y(u) = y_0 + u \Delta y_0 + \frac{u(u-1)}{2} \Delta^2 y(-1) + \frac{u(u^2-1)}{3!} \Delta^3 y(-1) + \dots \quad (9.3-4)$$

Можно выбирать все виды путей и каждый будет давать некоторую формулу. Нам нужно только доказать, что все это — действительно интерполяционные формулы. Для доказательства потребуется показать следующее:

1. Что получается по крайней мере одна правильная формула; так как из диаграммы мы нашли интерполяционную формулу Ньютона, то этот шаг доказательства уже сделан.

2. Что вклад по любому замкнутому пути равен нулю и, следовательно, что мы можем преобразовать один путь в любой другой.

3. Что если две формулы кончаются в одном и том же месте, то они одинаковы. Это необходимо доказывать, так как точки входа в ромбовидную диаграмму не обязаны быть одинаковыми для разных формул.

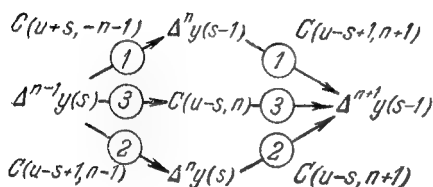


Рис. 9.3-2.

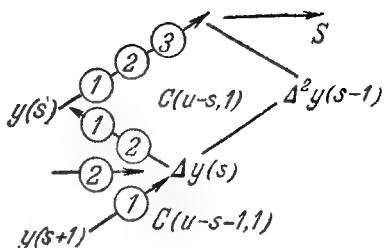


Рис. 9.3-3.

Чтобы доказать второе утверждение, возьмем ромб (рис. 9.3-2):

Путь 1: $C(u-s, n) \Delta^n y(s-1) +$
 $+ C(u-s+1, n+1) \Delta^{n+1} y(s-1).$

Путь 2: $C(u-s, n) \Delta^n y(s) + C(u-s, n+1) \Delta^{n+1} y(s-1).$

Путь 3: $C(u-s, n) \frac{\Delta^n y(s-1) + \Delta^n y(s)}{2} +$
 $+ \frac{C(u-s+1, n+1) + C(u-s, n+1)}{2} \Delta^{n+1} y(s-1).$

Вспоминая, что справа налево получаются отрицательные члены (правило 1б), мы должны лишь показать, что все три пути одинаковы. Но

$$\begin{aligned} \text{Путь 1} - \text{Путь 2} &= C(u-s, n) [\Delta^n y(s-1) - \Delta^n y(s)] + \\ &+ [C(u-s+1, n+1) - C(u-s, n+1)] \Delta^{n+1} y(s-1) = \\ &= C(u-s, n) [-\Delta^{n+1} y(s-1)] + \\ &+ C(u-s, n) \frac{u-s+1-(u-s-n)}{n+1} \Delta^{n+1} y(s-1) = \\ &= C(u-s, n) [-\Delta^{n+1} y(s-1) + \Delta^{n+1} y(s-1)] = 0. \end{aligned}$$

Кроме того, путь 3 равен среднему арифметическому путей 1 и 2. Таким образом, второй шаг доказательства закончен.

Чтобы провести третий шаг доказательства, используем рис. 9.3-3.

Путь 1: $y(s+1) + C(u-s-1, 1) \Delta y(s) - C(u-s, 1) \Delta y(s) + S.$

Путь 2: $\frac{y(s+1) + y(s)}{2} + \frac{C(u-s-1, 1) + C(u-s, 1)}{2} \Delta y(s) -$
 $- C(u-s, 1) \Delta y(s) + S.$

Путь 3: $y(s) + S.$

где S есть оставшаяся часть формулы. Теперь, используя равенство

$$y(s+1) - y(s) = \Delta y(s),$$

получаем

$$\begin{aligned} \text{Путь 1} = y(s) + \Delta y(s) + (n-s)\Delta y(s) - \Delta y(s) - \\ - (n-s)\Delta y(s) + S = y(s) + S = \text{Путь 3} \end{aligned}$$

Аналогично путь 2 можно свести к пути 3. Таким образом, мы можем сделать вывод, что интерполяционная формула зависит от окончательных значений и не зависит от пути, использованного для достижения их.

Упражнение 9.3-1. Проверить формулы (9.3-2) — (9.3-4).

§ 9.4. Замечания к выведенным формулам

Как сравнить эти разные формулы между собой и с формулой Лагранжа, найденной раньше? Значение, получаемое при интерполяции, зависит от используемого многочлена, а многочлен зависит от выбранных узловых точек. Остаточный член имеет форму (см. § 8.6)

$$\frac{(x-x_1)(x-x_2)\dots(x-x_{n-1})y^{(n+1)}(x)}{(n+1)!}.$$

Коэффициент при производной минимизируется, если x берется в середине интервала узловых точек. Поэтому, когда интерполируемая точка находится в середине интервала, обычно используют четное число узловых точек, а когда она лежит около узловой точки — нечетное.

Формулы, получаемые из ромбовидной диаграммы, содержат в явном виде разности, и эти разности дают некоторое представление о точности, но эти формулы требуют больше арифметических действий, чем метод интерполяции Лагранжа. Зато метод Лагранжа непосредственно не использует разности, так что нет показателя точности. Это — один из парадоксов численного анализа: использование разностей дает некоторые указания на точность примененного метода, в то время как метод Лагранжа минимизирует количество арифметических действий. В результате разностные методы обычно применяются при исследованиях, а метод Лагранжа — в обычной повседневной работе.

Коэффициенты Лагранжа для равноотстоящих узлов широко протабулированы [32]. Некоторые коэффициенты для кубической интерполяции даны в таблице 9.4-1. Заметим, что симметрия позволяет нам читать таблицу сверху, используя записи слева, или снизу, используя записи справа.

Т а б л и ц а 9.4-1

Кубическая интерполяция Лагранжа

$$f(x) = A_{-1}(x)f(-1) + A_0(x)f(0) + A_1(x)f(1) + A_2(x)f(2)$$

x	A_{-1}	A_0	A_1	A_2	
0	0	1	0	0	1,0
0,1	-0,0285	0,9405	0,1045	-0,0165	0,9
0,2	-0,0480	0,8640	0,2160	-0,0320	0,8
0,3	-0,0595	0,7735	0,3315	-0,0455	0,7
0,4	-0,0640	0,6720	0,4480	-0,0560	0,6
0,5	-0,0625	0,5625	0,5625	-0,0625	0,5
	A_2	A_1	A_0	A_{-1}	x

Упражнения. 9.4-1. Используя интерполяционные коэффициенты Лагранжа, вычислить интеграл ошибок в точке 1,40 по таблице 9.2-1.

9.4-2. Используя таблицу 9.4-1, вычислить значение в таблице для $x = 0,35$.

§ 9.5. Смешанные интерполяционные формулы

Иногда бывают полезны некоторые другие формулы, не получающиеся непосредственно из ромбовидной диаграммы. Они обычно опираются на то, что любая разность может быть исключена применением формулы

$$\Delta^n y(s+1) - \Delta^n y(s) = \Delta^{n+1} y(s).$$

Ценой этого исключения является включение дополнительной разности некоторого другого порядка.

Например, в интерполяционной формуле Бесселя можно исключить все разности нечетного порядка; это даст формулу

$$y(u) = (1-u)y(0) + uy(1) + \frac{(2-u)(1-u)(-u)}{3!} \Delta^2 y(-1) + \\ + \frac{(u+1)u(u-1)}{3!} \Delta^2 y(0) + \dots,$$

которая известна как формула Эверетта и очень популярна, так как создателю таблицы нужно опубликовать лишь функцию и разности четного порядка.

Аналогично можно начать почти с любой формулы и исключить разности любого порядка (ценою добавления разностей тех порядков, которые мы оставляем в формуле). Мы могли бы, если бы хотели, исключить, скажем, разности как второго, так и третьего порядков, используя лишь функцию, разности первого порядка, а также четвертые и более высокие разности. Проведение этой идеи до конца приведет к формуле Лагранжа, которая вообще не использует разностей,

но использует много значений функций. Исключение разностей производится главным образом для экономии знаков и места при печатании таблиц, и делается это за счет дополнительной работы со стороны потребителя. Надлежащее равновесие зависит от обстоятельств и не может быть достигнуто раз навсегда.

Второй прием, которым часто пользуются, называется «отбрасыванием». Здесь используется тот факт, что коэффициенты последовательных разностей в различных интерполяционных формулах, как, например, Эверетта, бывают пропорциональными друг другу, и поэтому, если высокие разности, взятые с соответствующими коэффициентами, при печатании таблиц объединить с низкими, то большая часть эффекта применения разностей высокого порядка в интерполировании автоматически достигается использованием формулы низкого порядка.

Так, в формуле Бесселя отношение коэффициента при Δ^4 к коэффициенту при Δ^2 равно

$$\frac{B^{IV}}{B^{II}} = \frac{(u+1)(u-2)}{12} \quad (0 \leq u \leq 1),$$

значение этого выражения лежит между $-\frac{1}{6}$ и $-\frac{3}{16}$. Следовательно, если мы добавим $C\Delta^4$ к Δ^2 и таким образом образуем столбец модифицированных вторых разностей, мы допустим ошибку $(B^{IV} - CB^{II})\Delta^4$. Число C часто берут равным $-0,184$, как среднее между $-\frac{1}{6}$ и $-\frac{3}{16}$.

Мы опять напоминаем читателю, что создание таблиц — это большое искусство, и отсылаем читателя к Фоксу [10] или к классическим руководствам Копала [20] и Хильдебранда [14].

Упражнения

9.5-1. Найти формулу, которая использует лишь функцию и третью и шестые разности, исключением других из интерполяционной формулы Бесселя.

9.5-2. То же, что в упражнении 9.5-1, но используя формулу Стирлинга.

ГЛАВА 10

ЕДИНЫЙ МЕТОД НАХОЖДЕНИЯ ИНТЕРПОЛЯЦИОННЫХ ФОРМУЛ

§ 10.1. Введение

Цель этой главы — дать единый метод для нахождения интерполяционных формул. В отношении четырех основных вопросов:

1. *Каковы узловые точки?*
2. *Каков класс функций?*
3. *Каков критерий согласия?*
4. *Какова точность?*

мы ограничимся в главах 10—16 классом многочленов и критерием точного совпадения. Преимущество предлагаемого метода в том, что он не только делает более легким изучение единого разностного метода, но и, что более важно, приводит к возможности вывода соответствующей формулы самой вычислительной машиной.

Поверхностные размышления о механизации численного анализа наводят на мысль, что все формулы могут быть зашифрованы каким-нибудь подходящим образом и расположены на запоминающей ленте машины. Потребитель должен лишь знать, какая формула ему нужна, и вызывать ее по ее шифру каждый раз, когда она понадобится. Я полагаю, что такой способ действия неудачен. Имеющийся опыт работы с обширной библиотекой подпрограмм, охватывающей огромное, плохо определенное поле человеческой мысли, показывает, что среди записанных на библиотечной ленте подпрограмм слишком часто не оказывается нужных.

Эту проблему можно рассматривать как задачу извлечения информации. Пусть мы хотим найти конкретный кусок информации, скажем, коэффициенты интерполяционной формулы седьмого порядка.

Вместо извлечения информации мы предлагаем использовать ее восстановление. Вместо того чтобы искать на ленте формулу, которая может быть, а может и не быть на ней, мы напишем программу, которая будет выводить любую формулу из широкого класса.

Этот прием не является совершенной новостью в вычислительной математике. Так, мы обычно заново вычисляем значения элементарных трансцендентных функций, вместо того чтобы справиться по таблице. Последние работы в области доказательств теорем наводят на мысль, что предлагаемый метод выведения формул заново всякий раз, когда мы нуждаемся в них, не такая уж невозможная задача, как могло показаться сначала.

Какой из двух методов, нахождение или восстановление информации, является наилучшим или, может быть, следует использовать их смесь — это отчасти вопрос экономический. Когда число формул, которые надо иметь, велико, то могут быть очень высокими цена хранения и время поиска, если же формула имеет длинную, сложную структуру, становится дорогостоящим время восстановления. Для ограниченного класса формул, который мы собираемся рассмотреть, оказывается, что метод восстановления вполне осуществим.

Предлагаемый метод имеет несколько достоинств. Во-первых, он является мощным стимулом для дальнейших исследований. Во-вторых, он таков, что потребитель может выбрать формулу, соответствующую ситуации, и не пытаться найти уже выведенную формулу, которая работает приближенно. В-третьих, он делает обучение много проще, так как применяется лишь один метод. Наконец, он в духе современной прикладной математики: распределяя работу между человеком и вычислительной машиной, мы хотим переложить сколько можно на

нее в той части работы, которую машина может сделать наилучшим образом.

В этой главе мы сосредоточимся на формальных методах вывода формул. В гл. 11 будут исследованы методы оценки ошибки, сделанной при применении формулы. В гл. 12 мы исследуем и сравним между собой несколько характерных формул для численного интегрирования. Читатель, которому нужна конкретная формула интегрирования, должен прежде всего поискать ее в других местах этой книги. В случае неудачи ему следует вернуться сюда и найти, как ее вывести. В главах 13 и 15 мы изложим развитие техники нахождения формул.

§ 10.2. Несколько типичных формул интегрирования

Прежде чем углубиться в изучение метода систематического выведения формул, рассмотрим несколько типичных примеров. На самом деле этот метод затрагивает много больше, чем только формулы интегрирования, но здесь проще говорить в терминах численного интегрирования. Можно подумать, что численное дифференцирование проще, чем интегрирование, но это не так. Задача приближенного вычисления производных из данных с погрешностями будет рассмотрена несколько позже.

В § 7.3 отмечалось, что задача численного интегрирования

$\int_0^1 f(x) dx$ эквивалентна оценке среднего арифметического значения $f(x)$

на отрезке $0 \leq x \leq 1$ по n узлам. Это — статистическая задача последовательной выборки и планирования экспериментов, и статистики начинают думать над такими проблемами. Однако, так как их результаты до сих пор скудны*), мы вернемся к классическим методам, которые заранее определяют как положения x_i , где должны быть взяты узлы, так и веса w_i , соответствующие этим узлам.

Не будем использовать аналитическую замену, т. е. построение интерполяционного многочлена, проходящего точно через узловые точки, и последующего интегрирования этого многочлена между пределами интегрирования; вместо этого найдем коэффициенты формулы непосредственно. Не будем также пользоваться методом рядов Тейлора, а сделаем формулу точной для полиномов возможно более высокой степени. В § 7.3 было показано, что эти три метода эквивалентны, а третий метод является самым легким для практического применения.

Простейшая нетривиальная формула для вычисления интеграла использует одну узловую точку, а именно:

$$\int_0^1 f(x) dx = w_1 f(x_1),$$

*) См. J. H. Halton, Thesis, Princeton University, Princeton, N. J.

где w_1 и x_1 подлежат определению. За счет выбора двух свободных параметров можно сделать формулу точной для $f(x)=1$ и $f(x)=x$. Используя эти две функции, получим уравнения

$$\int_0^1 1 dx = 1 = w_1, \quad \text{и} \quad \int_0^1 x dx = \frac{1}{2} = w_1 x_1,$$

откуда $w_1=1$, $x_1=\frac{1}{2}$. Таким образом, имеем формулу

$$\int_0^1 f(x) dx = f\left(\frac{1}{2}\right). \quad (10.2-1)$$

Разумно оценивать ошибку этой формулы, используя следующую более высокую степень x , в данном случае x^2 :

$$\int_0^1 x^2 dx = \frac{1}{3} = \frac{1}{4} + E_2 \quad \text{или} \quad E_2 = \frac{1}{12}.$$

В следующей главе будет показано, что ошибка, или остаточный член, как его часто называют, есть

$$\frac{h^3 E_2 f''(\theta)}{2!} = \frac{h^3 f''(\theta)}{24} = \frac{f''(\theta)}{24}, \quad (10.2-2)$$

где θ — некоторое число из интервала $(0,1)$ и h — длина интервала (в данном случае $h=1$).

Стандартная формула трапеций использует два узла в концах интервала:

$$\int_0^1 f(x) dx = w_0 f(0) + w_1 f(1).$$

Подставим опять $f(x)=1$ и $f(x)=x$

$$1 = w_0 + w_1, \quad \frac{1}{2} = w_1.$$

Из этих уравнений получаем

$$\int_0^1 f(x) dx = \frac{1}{2} [f(0) + f(1)]. \quad (10.2-3)$$

Чтобы оценить ошибку, возьмем опять функцию x^2 :

$$\frac{1}{3} = \frac{1}{2} + E_2 \quad \text{или} \quad E_2 = -\frac{1}{6}, \quad (10.2-4)$$

что вдвое больше ошибки формулы (10.2-2) с единственным узлом.

Можно поместить два узла в произвольные точки

$$\int_0^1 f(x) dx = w_1 f(x_1) + w_2 f(x_2). \quad (10.2-5)$$

Имея четыре свободных параметра, используем функции $1, x, x^2, x^3$:

$$\left. \begin{aligned} 1 &= w_1 + w_2, \\ \frac{1}{2} &= w_1 x_1 + w_2 x_2, \\ \frac{1}{3} &= w_1 x_1^2 + w_2 x_2^2, \\ \frac{1}{4} &= w_1 x_1^3 + w_2 x_2^3. \end{aligned} \right\} \quad (10.2-6)$$

Можно решить эти уравнения непосредственно, но проще перейти в другую систему координат, начало которой находится в середине отрезка интегрирования. Можно также изменить масштаб, не меняя условия, что формула должна быть точной для $1, x, x^2, x^3$. Поэтому преобразуем задачу так:

$$\int_{-1}^1 f(x) dx = w_1 f(x_1) + w_2 f(x_2), \quad (10.2-7)$$

что приводит к уравнениям:

$$\begin{aligned} 2 &= w_1 + w_2, \\ 0 &= w_1 x_1 + w_2 x_2, \\ \frac{2}{3} &= w_1 x_1^2 + w_2 x_2^2, \\ 0 &= w_1 x_1^3 + w_2 x_2^3. \end{aligned}$$

Теперь исключим $w_2 x_2$ из второго и четвертого уравнений и получим

$$0 = w_1 x_1^3 - w_1 x_1 x_2^2 \quad (w_1 x_1 \neq 0) *$$

или

$$x_1^2 = x_2^2, \quad x_1 = -x_2.$$

(Знак плюс дал бы лишь одну узловую точку.) Второе из четырех уравнений теперь дает $w_1 = w_2$, а первое уравнение дает $w_1 = w_2 = 1$. Наконец, третье уравнение дает

$$\frac{2}{3} = 2x_1^2 \quad \text{или} \quad x_1 = \frac{1}{\sqrt{3}}.$$

*) $w_1 x_1 = 0$ противоречит системе уравнений.

Формула, следовательно, такова:

$$\int_{-1}^1 f(x) dx = f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right).$$

Если положить $y = \frac{1+x}{2}$, то для первоначального интеграла (10.2-5) получаем

$$\begin{aligned} \int_0^1 f(y) dy &= \frac{1}{2} \left[f\left(\frac{1-\frac{1}{\sqrt{3}}}{2}\right) + f\left(\frac{1+\frac{1}{\sqrt{3}}}{2}\right) \right] = \\ &= \frac{1}{2} [f(0,2113) + f(0,7887 \dots)]. \end{aligned} \quad (10.2-8)$$

Такого типа формулы, в которых мы не фиксируем заранее узлы и веса, называются *гауссовыми*; впоследствии они будут изучены более детально. В качестве другого примера предположим, что известны как функция, так и ее производная в крайних точках отрезка. Тогда можно написать

$$\int_0^1 f(x) dx = w_0 f(0) + w_1 f(1) + w_2 f'(0) + w_3 f'(1)$$

и, подставляя $f(x) = 1, x, x^2, x^3$, получаем уравнения:

$$\begin{aligned} 1 &= w_0 + w_1, \\ \frac{1}{2} &= w_1 + w_2 + w_3, \\ \frac{1}{3} &= w_1 + 2w_3, \\ \frac{1}{4} &= w_1 + 3w_3, \end{aligned}$$

которые легко решаются:

$$w_3 = -\frac{1}{12}, \quad w_1 = \frac{1}{2}, \quad w_0 = \frac{1}{2}, \quad w_2 = \frac{1}{12}.$$

Имеем формулу

$$\int_0^1 f(x) dx = \frac{1}{2} [f(0) + f(1)] + \frac{1}{12} [f'(0) - f'(1)]. \quad (10.2-9)$$

Чтобы оценить ошибку, подставим $f(x) = x^4$:

$$\frac{1}{5} = w_1 + 4w_3 + E_4, \quad E_4 = \frac{1}{30}.$$

Используя результаты предыдущей главы, получаем ошибку

$$\frac{1}{30} \frac{h^5}{4!} f^{(IV)}(\theta) = \frac{1}{720} f^{(IV)}(\theta).$$

В качестве последнего примера, рассмотрим интеграл

$$\int_{-1}^1 f(x) \sin \frac{\pi}{2} x dx = w_{-1} f(-1) + w_0 f(0) + w_1 f(1). \quad (10.2-10)$$

Метод никоим образом не зависит от свойств функции $\sin x$ и так же хорошо приложим к интегралам вида

$$\int_a^b K(x) f(x) dx,$$

хотя веса w_i будут зависеть от выбора $K(x)$.

Имея три параметра, подставим 1, x , x^2 :

$$0 = w_{-1} + w_0 + w_1,$$

$$\frac{8}{\pi^2} = -w_{-1} + w_1,$$

$$0 = w_{-1} + w_1.$$

Полученные уравнения легко решаются:

$$w_1 = -w_{-1} = \frac{4}{\pi^2}, \quad w_0 = 0,$$

что дает для (10.2-10)

$$\int_{-1}^1 f(x) \sin \frac{\pi}{2} x dx = \frac{4}{\pi^2} [f(1) - f(-1)]. \quad (10.2-11)$$

Упражнения

10.2-1. Получить формулу

$$\int_{-\pi}^{\pi} f(x) \sin x dx = \left(1 - \frac{8}{\pi^2}\right) [f(\pi) - f(-\pi)] + \frac{16}{\pi^2} \left[f\left(\frac{\pi}{2}\right) - f\left(-\frac{\pi}{2}\right)\right].$$

Матрица неизвестных $X = (x_i^k)$ имеет определитель (определитель Вандермонда), который отличен от нуля, если $x_i \neq x_j$ (см. § 8.2).

Мы намерены исследовать аналитическое обращение матрицы X . Можно, конечно, обращаться матрицу каждый раз численно, но если бы это делалось, то возникли бы вопросы точности, на которые трудно ответить; действительно, в конкретных задачах часто бывает трудно обратить матрицу. Если же дана аналитическая форма обратной матрицы, то относительно легко оценить точность вычисления. Чтобы найти обратную матрицу, введем фундаментальные многочлены (ср. § 8.3)

$$\begin{aligned}\pi_i(x) &= (x - x_1)(x - x_2) \dots (x - x_{i-1})(x - x_{i+1}) \dots (x - x_n) = \\ &= \sum_{k=0}^{n-1} C_{i,k} x^k \quad (i = 1, 2, \dots, n), \quad (10.3-2)\end{aligned}$$

где i -й многочлен не содержит i -го множителя $(x - x_i)$. Отметим тот важный факт, что

$$\pi_i(x_j) = 0 \quad (i \neq j) \quad \text{и} \quad \pi_i(x_i) \neq 0.$$

Это обстоятельство подсказывает способ обращения X . Так как $X^{-1}X = 1$, то первая строка X^{-1} должна состоять из элементов вида $\frac{C_{1,k}}{\pi_1(x_1)}$. Действительно, умножение на j -й столбец X дает

$$\frac{\sum C_{1,k} x_j^k}{\pi_1(x_1)} = \frac{\pi_1(x_j)}{\pi_1(x_1)} = \begin{cases} 1, & j = 1, \\ 0, & j \neq 1. \end{cases}$$

Вообще m -я строка X^{-1} может быть написана в виде

$$\frac{C_{m,k}}{\pi_m(x_m)},$$

где $C_{m,k}$ — симметричные функции соответствующих узлов, которые легко находятся при помощи умножений и сложений; они являются многочленами относительно узлов x_i .

Частный случай равноотстоящих узлов, когда они симметрично расположены около начала координат, стоит затабулировать. Обратные матрицы для этого случая обозначены через S_n и протабулированы ниже. Числа в скобках, следующих за S , указывают положение узлов, так что $S_4\left(-\frac{3}{2}, -\frac{1}{2}, \frac{1}{2}, \frac{3}{2}\right)$ означает, что узлы находятся в $-\frac{3}{2}, -\frac{1}{2}, \frac{1}{2}, \frac{3}{2}$. Если матрицу умножить справа на столбец вектора моментов, вектор-столбец результата даст веса.

Универсальные матрицы

$$S_2 = S_2\left(-\frac{1}{2}, \frac{1}{2}\right) = \begin{pmatrix} \frac{1}{2} & -1 \\ \frac{1}{2} & 1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & -2 \\ 1 & 2 \end{pmatrix},$$

$$S_3 = S_3(-1, 0, 1) = \begin{pmatrix} 0 & -\frac{1}{2} & \frac{1}{2} \\ 1 & 0 & -1 \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 0 & -1 & 1 \\ 2 & 0 & -2 \\ 0 & 1 & 1 \end{pmatrix},$$

$$S_4 = S_4\left(-\frac{3}{2}, -\frac{1}{2}, \frac{1}{2}, \frac{3}{2}\right) = \frac{1}{48} \begin{pmatrix} -3 & 2 & 12 & -8 \\ 27 & -54 & -12 & 24 \\ 27 & 54 & -12 & -24 \\ -3 & -2 & 12 & 8 \end{pmatrix},$$

$$S_5 = S_5(-2, -1, 0, 1, 2) = \frac{1}{24} \begin{pmatrix} 0 & 2 & -1 & -2 & 1 \\ 0 & -16 & 16 & 4 & -4 \\ 24 & 0 & -30 & 0 & 6 \\ 0 & 16 & 16 & -4 & -4 \\ 0 & -2 & -1 & 2 & 1 \end{pmatrix},$$

$$S_6 = S_6\left(-\frac{5}{2}, -\frac{3}{2}, \dots, \frac{5}{2}\right) =$$

$$= \frac{1}{3840} \begin{pmatrix} 45 & -18 & -200 & 80 & 80 & -32 \\ -375 & 250 & 1560 & -1040 & -240 & 160 \\ 2250 & -4500 & -1360 & 2720 & 160 & -320 \\ 2250 & 4500 & -1360 & -2720 & 160 & 320 \\ -375 & -250 & 1560 & 1040 & -240 & -160 \\ 45 & 18 & -200 & -80 & 80 & 32 \end{pmatrix},$$

$$S_7 = \frac{1}{720} \begin{pmatrix} 0 & -12 & 4 & 15 & -5 & -3 & 1 \\ 0 & 108 & -54 & -120 & 60 & 12 & -6 \\ 0 & -540 & 540 & 195 & -195 & -15 & 15 \\ 720 & 0 & -980 & 0 & 280 & 0 & -20 \\ 0 & 540 & 540 & -195 & -195 & 15 & 15 \\ 0 & -108 & -54 & 120 & 60 & -12 & -6 \\ 0 & 12 & 4 & -15 & -5 & 3 & 1 \end{pmatrix}.$$

§ 10.4. Некоторые примеры формул

Исследуем разновидности формул, которые можно получить из этих матриц. Рассмотрим $S_3 = S_3(-1, 0, 1)$:

$$S_3 = \begin{pmatrix} 0 & -\frac{1}{2} & \frac{1}{2} \\ 1 & 0 & -1 \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Предположим, требуется вывести формулу Симпсона для интегрирования

$$\int_{-1}^1 f(x) dx = w_{-1}f(-1) + w_0f(0) + w_1f(1).$$

Нужно найти моменты m_0, m_1, m_2 , которые определяются как

$$m_k = \int_{-1}^1 x^k dx = \frac{1 + (-1)^k}{k+1}.$$

Считая их вектором-столбцом m , умножим его слева на S_3 , чтобы получить веса w_i в виде вектора-столбца w :

$$S_3 m = \begin{pmatrix} 0 & -\frac{1}{2} & \frac{1}{2} \\ 1 & 0 & -1 \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 2 \\ 0 \\ \frac{2}{3} \end{pmatrix} = \begin{pmatrix} \frac{1}{3} \\ \frac{4}{3} \\ \frac{1}{3} \end{pmatrix} = w, \quad (10.4-1)$$

что, как известно, является правильным ответом (см. (7.2-4)).

Далее, рассмотрим нахождение «половинной формулы Симпсона» для

$$\int_{-1}^0 f(x) dx = a_{-1}f(-1) + a_0f(0) + a_1f(1).$$

Моменты равны

$$m_0 = 1, \quad m_1 = -\frac{1}{2}, \quad m_2 = \frac{1}{3}.$$

Следовательно,

$$\begin{pmatrix} 0 & -\frac{1}{2} & \frac{1}{2} \\ 1 & 0 & -1 \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 \\ -\frac{1}{2} \\ \frac{1}{3} \end{pmatrix} = \begin{pmatrix} \frac{5}{12} \\ \frac{8}{12} \\ -\frac{1}{12} \end{pmatrix} = \frac{1}{12} \begin{pmatrix} 5 \\ 8 \\ -1 \end{pmatrix}. \quad (10.4-2)$$

Попробуем теперь найти формулу вида

$$\frac{dy}{dx}\bigg|_{x=0} = w_{-1}f(-1) + w_0f(0) + w_1f(1).$$

Вектор моментов здесь

$$\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

и мы имеем

$$\begin{pmatrix} 0 & -\frac{1}{2} & \frac{1}{2} \\ 1 & 0 & -1 \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -\frac{1}{2} \\ 0 \\ \frac{1}{2} \end{pmatrix}. \quad (10.4-3)$$

Последним примером § 10.2 (уравнение (10.2-10)) была формула для интеграла

$$\int_{-1}^1 f(x) \sin \frac{\pi}{2} x dx,$$

моменты которого 0 , $\frac{8}{\pi^2}$ и 0 . Следовательно, имеем

$$\begin{pmatrix} 0 & -\frac{1}{2} & \frac{1}{2} \\ 1 & 0 & -1 \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 0 \\ \frac{8}{\pi^2} \\ 0 \end{pmatrix} = \begin{pmatrix} -\frac{4}{\pi^2} \\ 0 \\ \frac{4}{\pi^2} \end{pmatrix},$$

что совпадает с (10.2-11).

Предположим, что нужно проинтерполировать в точке x , используя те же самые три узла. На этот раз вектор моментов зависит от положения точки x , в которой мы интерполируем. Получаем

$$\begin{pmatrix} 0 & -\frac{1}{2} & \frac{1}{2} \\ 1 & 0 & -1 \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 \\ x \\ x^2 \end{pmatrix} = \begin{pmatrix} \frac{x^2 - x}{2} \\ 1 - x^2 \\ \frac{x^2 + x}{2} \end{pmatrix} \quad (10.4-4)$$

или обычное:

$$f(x) = \frac{x^2 - x}{2} f(-1) + (1 - x^2) f(0) + \frac{x^2 + x}{2} f(1),$$

что, конечно, является правильным ответом.

Этот последний пример наводит на мысль, что можно найти формулы для неопределенных интегралов так же, как и для определенных. Если подумать еще немного, становится ясно, что такие интегралы, как

$$\int_{-1}^0 f(x) dx, \quad \int_0^5 f(x) dx, \quad \int_0^x K(x) f(x) dx,$$

могут быть найдены в том же самом виде при условии, что мы можем найти необходимые моменты.

Упражнения

10.4-1. Получить результат упражнения 10.2-1.

10.4-2. Найти интерполяционный многочлен четвертой степени, соответствующий (10.4-4).

10.4-3. Найти формулы для $y'(0)$, $y''(0)$, $y'''(0)$, $y^{IV}(0)$, используя S_R .

§ 10.5. Значения функции и производной в фиксированных точках

Часто случается, что в узловых точках мы либо уже имеем как значения функции, так и значения производных, либо, найдя значения функции, можем, затратив лишь небольшую дополнительную работу, найти производные. Хотя можно использовать несколько производных, ограничим наше рассмотрение случаем одной производной; к более сложным случаям легко перейти, если понята основная идея вывода таких формул. Для стандартизации поместим сначала все значения функции, а следом за ними значения производной, как в (10.2-9). Несмотря на справедливость общего правила «никогда не игнорировать любую информацию, которую вам случится иметь», иногда бывает слишком дорого использовать все, что имеешь.

Итак, правая часть формулы имеет вид

$$w_1 f(x_1) + w_2 f(x_2) + \dots + w_n f(x_n) + w_{n+1} f'(x_1) + \dots + w_{2n} f'(x_n).$$

Таким образом, подлежащая обращению матрица есть

$$\begin{pmatrix} 1 & 1 & \dots & 1 & 0 & & 0 & \dots & 0 \\ x_1 & x_2 & \dots & x_n & 1 & & 1 & \dots & 1 \\ x_1^2 & x_2^2 & \dots & x_n^2 & 2x_1 & & 2x_2 & \dots & 2x_n \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ x_1^{2n-1} & x_2^{2n-1} & \dots & x_n^{2n-1} & (2n-1)x_1^{2n-2} & (2n-1)x_2^{2n-2} & \dots & (2n-1)x_n^{2n-2} \end{pmatrix}.$$

На этот раз мы хотим найти коэффициенты многочлена, удовлетворяющего всем уравнениям

$$\begin{aligned} P(x_i) &= 0, \\ P'(x_i) &= 0, \end{aligned} \quad (i = 1, 2, \dots, n),$$

Упражнения

10.5-1. Показать, что обратная матрица для случая, когда значения функции и первой производной заданы в точках $x = -1, 0, 1$, есть

$$\begin{pmatrix} 0 & 0 & 1/4 & -5/4 & -2/4 & 8/4 \\ 1 & 0 & -2 & 0 & 1 & 0 \\ 0 & 0 & 1/4 & 5/4 & -2/4 & -8/4 \\ 0 & 0 & 1/4 & -1/4 & -1/4 & 1/4 \\ 0 & 1 & 0 & -2 & 0 & 1 \\ 0 & 0 & -1/4 & -1/4 & 1/4 & 1/4 \end{pmatrix}.$$

10.5-2. Применить результат упражнения 10.5-1 к интегралу в упражнении 10.2-2.

§ 10.6. Свободные узлы; квадратура Гаусса

В § 10.2 исследовалась пара формул (10.2-1) и (10.2-8), в которых положения всех узлов, так же как и все веса, считались параметрами. Общая формула такого типа

$$\int_a^b k(x) f(x) dx = \sum_{i=1}^n w_i f(x_i), \quad k(x) > 0, \quad (10.6-1)$$

имеет $2n$ параметров и может быть сделана точной для $f(x) = 1, x, \dots, x^{2n-1}$. Мы получаем определяющие уравнения

$$\left. \begin{aligned} m_0 &= w_1 & + w_2 & + w_3 & + \dots & + w_n \\ m_1 &= w_1 x_1 & + w_2 x_2 & + w_3 x_3 & + \dots & + w_n x_n \\ m_2 &= w_1 x_1^2 & + w_2 x_2^2 & + w_3 x_3^2 & + \dots & + w_n x_n^2 \\ &\dots & \dots & \dots & \dots & \dots \\ m_{2n-1} &= w_1 x_1^{2n-1} & + w_2 x_2^{2n-1} & + \dots & & + w_n x_n^{2n-1} \end{aligned} \right\} \quad (10.6-2)$$

Существует хорошо известный метод для решения этой частной системы нелинейных уравнений. Сначала определим узловой многочлен

$$\pi(x) = \prod_{i=1}^n (x - x_i) = \sum_{k=0}^n C_k x^k. \quad (10.6-3)$$

Заметим, что

$$\pi(x_i) = 0 \quad (i = 1, 2, \dots, n).$$

Умножив первое из уравнений (10.6-2) на C_0 следующее на C_1 и т. д., n -е на $C_n = 1$ и сложив, получим

$$\sum_{k=0}^n C_k m_k = \sum_{i=1}^n w_i \pi(x_i) = 0.$$

Теперь сдвинем каждый коэффициент вниз на одно уравнение и, повторив процесс, найдем

$$\sum_{k=0}^n C_k m_{k+1} = \sum_{i=1}^n w_i x_i \pi(x_i) = 0.$$

Если проделать все это n раз, то получится система линейных уравнений относительно C_k ($k = 0, 1, \dots, n-1$)

$$\sum_{k=0}^n C_k m_{k+j} = 0 \quad (j = 0, 1, \dots, n-1). \quad (10.6-4)$$

Определитель назван «персимметричным» (см. [30], стр. 419). Если $\|m_{k+j}\| \neq 0$, то можно решить систему относительно C_k (приняв $C_n = 1$).

Если определитель равен нулю, то может случиться одно из двух. Во-первых, уравнения могут быть несовместными. Пример:

$$\int_{-1}^1 x f(x) dx = w_1 f(x_1),$$

$$m_0 = 0 = w_1 \cdot 1, \quad m_1 = \frac{2}{3} = w_1 \cdot 0 \text{ — противоречие.}$$

Следовательно, такой формулы нет.

Во-вторых, ранг матрицы может оказаться равным $n-1$. Тогда в группу определяющих уравнений (10.6-2) добавляем еще одно и выводим еще одно уравнение (10.6-4), включающее m_{k+j} и C_k . Этот процесс повторяется до тех пор, пока не получится n линейно независимых уравнений для C_k , которые могут быть, а могут и не быть совместными. Каждое дополнительное уравнение указывает на повышенную точность формулы. По коэффициентам C_k узлового многочлена (10.6-3)

$$\pi(x) = x^n + C_{n-1}x^{n-1} + \dots + C_0 = 0$$

можно найти нули $\pi(x)$, которые являются узлами x_i .

Теперь все свелось к предыдущему случаю с известными узлами и можно использовать первые n определяющих уравнений (10.6-2), чтобы найти веса w_i .

В качестве примера рассмотрим уравнения (10.2-6), которые мы не стали решать в § 10.2 из-за громоздкости выкладок. Там получили (10.2-5) и (10.2-6):

$$\int_0^1 f(x) dx = w_1 f(x_1) + w_2 f(x_2),$$

$$\begin{array}{l} 1 = w_1 + w_2 \\ \frac{1}{2} = w_1 x_1 + w_2 x_2 \\ \frac{1}{3} = w_1 x_1^2 + w_2 x_2^2 \\ \frac{1}{4} = w_1 x_1^3 + w_2 x_2^3 \end{array} \quad \left| \begin{array}{l} C_0 \\ C_1 \\ 1 \\ 1 \end{array} \right| \quad \begin{array}{l} C_0 \\ C_1 \\ C_1 \\ 1 \end{array}$$

$$\pi(x) = (x - x_1)(x - x_2) = x^2 + C_1 x + C_0.$$

Множители для образования новых уравнений указаны во втором столбце справа. Используя их, получаем

$$C_0 + \left(\frac{1}{2}\right) C_1 + \frac{1}{3} = 0, \quad \left(\frac{1}{2}\right) C_0 + \left(\frac{1}{3}\right) C_1 + \frac{1}{4} = 0,$$

откуда

$$C_0 = \frac{1}{6}, \quad C_1 = -1.$$

Следовательно,

$$\pi(x) = x^2 - x + \frac{1}{6}$$

и

$$x_1 = \frac{1 - \sqrt{1 - \frac{2}{3}}}{2} = \frac{1 - \sqrt{\frac{1}{3}}}{2}, \quad x_2 = \frac{1 + \sqrt{1 - \frac{2}{3}}}{2} = \frac{1 + \sqrt{\frac{1}{3}}}{2},$$

как и раньше (см. (10.2-8)).

§ 10.7. Смешанный случай

Часто случается, что в качестве узловых мы хотим использовать одну или обе концевые точки. Рассмотрим использование лишь одной из концевых точек. При этом, конечно, теряется один из $2n$ параметров. Определяющие уравнения (10.3-1) запишем в виде

$$m_k = w_1 a^k + \sum_{i=2}^n w_i x_i^k \quad (k = 0, 1, \dots, 2n - 2),$$

где a — данная узловая точка.

Частично исключим a , умножая каждое уравнение на a и вычитая его из следующего уравнения:

$$\bar{m}_k = m_{k+1} - am_k = \sum_{i=2}^n w_i (x_i - a) x_i^k \quad (k = 0, 1, \dots, 2n - 3).$$

Теперь образуем узловой многочлен, используя лишь неизвестные узловые точки

$$\pi(x) = \prod_{i=2}^n (x - x_i) = \sum_{k=0}^{n-1} C_k x^k,$$

и повторим процесс исключения, примененный в предыдущем разделе:

$$\sum_{k=0}^{n-1} C_k m_{k+j} = \sum_{i=2}^n w_i (x_i - a) \pi(x_i) = 0 \quad (j = 0, 1, \dots, n - 2).$$

После этого можно найти C_k и неизвестные x_i , как и раньше.

Упражнения

10.7-1. Рассмотреть случай, когда в формуле должны быть использованы обе концевые точки.

10.7-2. Рассмотреть общую задачу исключения любого числа фиксированных узлов.

§ 10.8. Замечания

В этом параграфе будут рассмотрены попытки применить разобранные выше методы. Таким образом, он не является частью программы для нахождения формулы, но является исследованием того, что получится в процессе отыскания формулы.

Нет сомнения, что можно перейти от определяющих уравнений (10.6-2) к уравнениям (10.6-4)

$$\sum_{k=0}^n C_k m_{k+j} = 0.$$

Если ранг матрицы (m_{k+j}) недостаточно высок, чтобы позволить найти C_k , то мы добавляем еще одно определяющее уравнение и опять пробуем решить систему и т. д. В конце концов, мы либо повысим ранг до требуемой величины и решим систему, либо получим несовместные уравнения, показывающие, что такой формулы нет (либо нам надоест). Каждый раз добавляется еще одно определяющее уравнение и получается формула, которая является точной для следующей степени x ; маловероятно, кроме самых тривиальных случаев, что процесс не остановится очень быстро.

Следующая трудность возникает при попытке найти корни узловых многочленов. Покажем, что во многих случаях отыскиваемые

узлы действительны, различны и лежат в интервале интегрирования.

Прежде всего, рассмотрим случай, когда узлы не фиксированы и $K(x) \geq 0$. Так как формула точна для всех степеней x вплоть до $2n-1$, то имеем $[\pi(x)$ здесь то же, что и в уравнении (10.6-3)]

$$\int_a^b K(x) \pi(x) x^k dx = \sum_{i=1}^n w_i x_i^k \pi(x_i) = 0 \quad (k=0, \dots, n-1) \quad (10.8-1)$$

(так как все значения $\pi(x_i)$, встречающиеся справа, равны нулю). Теперь предположим, что $\pi(x)$ имеет меньше, чем n различных нулей в области интегрирования. Пусть

$$x_1, x_2, \dots, x_m \quad (m < n)$$

будут нули нечетной кратности на отрезке интегрирования. Многочлен

$$p(x) = \prod_{i=1}^m (x - x_i)$$

имеет степень $m \leq n-1$, следовательно, по (10.8-1) получаем

$$\int_a^b K(x) \pi(x) p(x) dx = 0.$$

Но в интервале интегрирования подынтегральная функция имеет постоянный знак и не может быть тождественным нулем. Следовательно, мы пришли к противоречию. Таким образом, в интервале интегрирования есть n действительных различных нулей $\pi(x)$ и мы гарантированы, что искомые действительные корни в самом деле существуют.

В случае, когда фиксирована одна концевая точка a , мы имеем узловой многочлен $\pi(x)$ степени $n-1$. образуем

$$\int_a^b K(x) (x-a) \pi(x) x^k dx = 0 \quad (k=0, 1, \dots, n-2)$$

и допустим, что $\pi(x)$ имеет лишь $m < n-1$ различных корней нечетного порядка в интервале. Построив, как и раньше, произведение $p(x)$, получаем

$$\int_a^b K(x) (x-a) \pi(x) p(x) dx = 0.$$

Это приводит к противоречию.

В качестве упражнения предоставляем читателю показать, что аналогичные соображения применимы и в случае, когда фиксированы обе концевые точки.

Можно также показать, что веса w_i в свободных узлах положительны. Это — интересный результат; он показывает, что формула будет стремиться сопротивляться погрешностям в исходных данных, так как не будет взаимных уничтожений из-за весов с противоположными знаками (хотя возможны уничтожения из-за функции, имеющей противоположные знаки в различных узлах). Для гауссова случая свободных узлов узловой многочлен определяется равенством (ср. (10.3-2))

$$\pi_k(x) = (x - x_1)(x - x_2) \dots (x - x_{k-1})(x - x_{k+1}) \dots (x - x_n).$$

Так как $\pi_k^2(x)$ имеет степень $2n - 2$, то

$$\int_a^b K(x) \pi_k^2(x) dx = \sum_{i=1}^n w_i \pi_k^2(x_i) = w_k \pi_k^2(x_k).$$

Следовательно $w_k > 0$ при условии, что $K(x) \geq 0$. Подобный же прием применяется в случаях, когда используются одна или обе концевые точки. Когда фиксированный узел лежит внутри интервала, этот результат, вообще говоря, неверен, хотя могут быть найдены частные случаи, когда он верен.

Не удастся показать, что веса положительны, когда фиксированы все узлы, потому что в действительности это неверно: они могут быть и отрицательны.

Если попробовать обобщить формулу гауссова типа, используя значения функции и ее первой производной в тех же узлах, то мы найдем (по крайней мере в нескольких случаях, которые исследованы), что нули узлового многочлена — комплексные числа.

Упражнение 10.8-1. Исследовать случай, когда фиксированы обе концевые точки. Показать, что все нули действительны, различны и находятся в интервале интегрирования и что все веса положительны.

§ 10.9. Линейные ограничения на веса

Иногда случается, что нам даны или желательны линейные ограничения на веса. Такие ограничения, вероятно, могут связывать веса, соответствующие фиксированным узлам, но они вряд ли могут налагаться на веса, соответствующие свободным узлам: трудно знать, в каком порядке они будут обозначены, — упорядочить их по величине является трудной математической задачей. Главное исключение: когда одно и то же ограничение применяется ко всем известным весам.

Хорошо известный пример ограничений на веса дает интегрирование по Чебышеву, где

$$w_1 = w_2 = \dots = w_n = w$$

и все узлы свободны. Это дает $n+1$ параметр и соответственно $n+1$ определяющее уравнение

$$m_k = w \sum_{i=1}^n x_i^k \quad (k=0, 1, \dots, n).$$

Полагая $k=0$, находим

$$w = \frac{m_0}{n}.$$

Остальные уравнения принимают вид

$$\frac{nm_k}{m_0} = \bar{m}_k = \sum_{i=1}^n x_i^k.$$

Тождества Ньютона позволяют выразить суммы степеней корней через коэффициенты C_k узлового многочлена. Отсюда получается узловой многочлен, что позволяет найти нули x_i .

Известно, что, когда $K(x)=1$, все узлы действительны, различны и находятся в интервале интегрирования для $n=1, 2, \dots, 9$ и только для этих n .

Ральстон *) опубликовал интересный пример смешанного типа интегрирования, в котором фиксированы два узла и одна линейная связь. Именно, формула

$$\int_{-1}^1 f(x) dx = \sum_{i=1}^n w_i f(x_i)$$

подчинена условиям:

$$x_1 = -1, \quad x_n = 1, \quad w_1 = -w_n.$$

Здесь имеется, следовательно, $2n-3$ параметра и формулу можно сделать точной для $1, x, \dots, x^{2n-4}$. Определяющие уравнения суть

$$\begin{aligned} 2 &= \sum_{i=2}^{n-1} w_i, \\ 0 &= -2w_1 + \sum w_i x_i, \\ \frac{2}{3} &= \sum w_i x_i^2, \\ 0 &= -2w_1 + \sum w_i x_i^3, \\ &\dots \dots \dots \\ \frac{2}{2n-8} &= \sum w_i x_i^{2n-4}. \end{aligned}$$

*) A. Ralston, J. Assoc. Computing Machinery, vol. 6, pp. 384—394, July 1959.

Чтобы сделать последующую выкладку более легкой и в то же время не вполне тривиальной, возьмем случай $n=5$, который дает $2n-3=7$ уравнений. Узловой многочлен есть

$$\pi(x) = (x-x_2)(x-x_3)(x-x_4) = x^3 + C_2x^2 + C_1x + C_0.$$

Обычный процесс исключения приводит к $n-1$ (в этом случае, четырем) уравнению:

$$\begin{aligned} C_0 + \frac{1}{3}C_2 &= -w_1C_1 - w_1, \\ \frac{1}{3}C_1 + \frac{1}{5} &= -w_1C_0 - w_1C_2, \\ \frac{1}{3}C_0 + \frac{1}{5}C_2 &= -w_1C_1 - w_1, \\ \frac{1}{5}C_1 + \frac{1}{7} &= -w_1C_0 - w_1C_2. \end{aligned}$$

Легко видеть, что можно образовать $n-3$ ($=2$) уравнения, которые не включают членов второго порядка w_1C_l ($l=0, 1, 2$),

$$\frac{2}{3}C_0 + \frac{2}{15}C_2 = 0, \quad \frac{2}{15}C_1 + \frac{2}{35} = 0.$$

Имеется, следовательно, $n-3$ линейных неоднородных уравнения с неизвестными C_0, C_1, \dots, C_{n-3} или на одно неизвестное больше, чем уравнений. Решая уравнения в зависимости, скажем, от C_0 , получаем

$$C_1 = -\frac{3}{7}, \quad C_2 = -5C_0.$$

Подставив их в первые два уравнения системы, получим

$$-\frac{1}{3}C_0 = -\frac{2}{7}w_1, \quad \frac{1}{35} = +2C_0w_1.$$

В общем случае, так же как при $n=5$, исключая C_0 , приходим к квадратному уравнению для w_1 . В случае $n=5$

$$w_1 = \pm \sqrt{\frac{1}{60}}.$$

Выбор знака w_1 определяет одну из двух симметричных формул. Зная w_1 , можно легко найти C_1 и узловой многочлен

$$\pi(x) = x^3 - 5\sqrt{\frac{3}{245}}x^2 - \frac{3}{7}x + \sqrt{\frac{3}{245}}.$$

Такие же соображения, как в § 10.8, показывают, что нули действительны, различны и находятся в интервале интегрирования.

Упражнение 10.9-1. Показать, что случаю Ральстона при $n=6$ соответствует узловой многочлен

$$\pi(x) = x^4 + \frac{4}{3\sqrt{6}}x^3 - \frac{2}{3}x^2 - \frac{4}{7\sqrt{6}}x + \frac{1}{21}.$$

§ 10.10. Формула Грегори

В обоих рассмотренных случаях, содержащих линейные ограничения на коэффициенты, большинство узлов было свободно. В этом параграфе рассматривается случай равноотстоящих фиксированных узлов со многими связями. Вводится новая концепция формулы, имеющей произвольное число узлов и определенное общее расположение весов. Для определенности исследуем формулы, веса которых расположены так:

$$a \ b \ b \ b \ b \ \dots \ b \ b \ b \ a, \quad (10.10-1)$$

$$a \ b \ c \ c \ c \ \dots \ c \ c \ b \ a, \quad (10.10-2)$$

$$a \ b \ c \ d \ d \ \dots \ d \ c \ b \ a \quad (10.10-3)$$

и т. д. Так как каждая формула имеет на один параметр больше, чем предыдущая, то можем надеяться найти семейство формул с возрастающей точностью. Формула (10.10-1), например, предполагает, что с использованием $n+1$ узла интеграл

$$\int_0^{nh} f(x) dx = af(0) + bf(h) + bf(2h) + \dots + bf[(n-1)h] + af(nh)$$

может быть найден точно как для $f(x)=1$, так и для $f(x)=x$. Это приводит к двум уравнениям (используя равенство (3.1-3)):

$$nh = b(n-1) + 2a$$

и

$$\frac{n^2 h^2}{2} = bh \frac{n(n-1)}{2} + ahn,$$

которое совпадает с первым. Подставив теперь $f(x)=x^2$, получаем

$$\frac{n^3 h^3}{3} = bh^3 \frac{n(n-1)(2n-1)}{6} + ah^3 n^2.$$

Решение этих двух уравнений дает

$$a = \frac{n}{2(n+1)} h = \frac{h}{2} - \frac{h}{2(n+1)}, \quad b = \frac{n^2}{n^2-1} h = h + \frac{h}{n^2-1}.$$

Таким образом, формула является точной для многочлена второй степени. Коэффициенты зависят от числа интервалов n , а это часто является неудобством. Можно потребовать, чтобы первое выведенное

уравнение было верно для всех n , и не пытаться сделать формулу точной для x^2 . Этот подход дает

$$b = h, \quad a = \frac{h}{2},$$

т. е. известное правило трапеций.

Схемы расположения весов выбраны симметричными; это означает, что если какая-нибудь из таких формул является точной для всех степеней x вплоть до некоторой четной, то она является точной для следующей более высокой нечетной степени. Этот эффект уже наблюдался в предыдущем примере, когда и 1, и x давали одно и то же уравнение. Чтобы убедиться в справедливости этого замечания в общем случае, заметим, что преобразование $\bar{x} = x - \frac{n}{2}$ переводит начало координат в середину области интегрирования. После этого все нечетные степени автоматически интегрируются точно, так как обе части уравнения вследствие симметрии обращаются в нуль.

При обратном преобразовании начала нечетная степень дает и все более низкие степени.

Если формула является точной для этих более низких степеней, то для них обе части равенства уничтожаются и формула оказывается верной для оставшейся старшей нечетной степени. Остается рассмотреть только четные степени x , так как симметричность задачи автоматически обеспечивает точность для нечетных степеней.

Постараемся найти формулы, в которых схемы расположения весов не зависят от количества взятых узлов, при условии, конечно, что их настолько много, чтобы появились все параметры. Эти формулы могли бы быть найдены раз навсегда независимо друг от друга, но одно простое замечание сводит задачу к тому, чтобы, преобразовывая каждую формулу по очереди, получать следующую из этого семейства.

Третья схема (10.10-3), например, может быть записана в виде

$$\int_0^{nh} f(x) dx = (\text{правило трапеций}) + A_1 [\Delta f(0) - \Delta f(n-1)] + \\ + A_2 [\Delta^2 f(0) + \Delta^2 f(n-2)].$$

Добавление разностных членов не повлияет на точность формулы для 1 и x . Очевидно также, что существуют такие константы A_1 и A_2 , что формула сведется к схеме (10.10-3).

В общем случае формула

$$\int_0^{nh} f(x) dx = (\text{правило трапеций}) + A_1 [\Delta f(0) - \Delta f(n-1)] + \\ + A_2 [\Delta^2 f(0) + \Delta^2 f(n-2)] + A_3 [\Delta^3 f(0) - \Delta^3 f(n-3)] + \\ + A_4 [\Delta^4 f(0) + \Delta^4 f(n-4)] + \dots$$

дает метод, при котором все коэффициенты могут быть, очевидно, найдены по очереди. В самом деле, пара коэффициентов определяется одновременно вследствие симметрии, которая обеспечивает точность формулы для нечетной степени. Таким образом, чтобы определить A_1 и A_2 , потребуем, чтобы формула была точной для x^3 . Используя результаты § 3.1, имеем (для $h=1$)

$$\frac{n^3}{3} = \frac{n(n-1)(2n-1)}{6} + \frac{1}{2}n^2 + A_1[1 - (2n-1)] + A_2[2 + 2].$$

Приравнявая коэффициенты при одинаковых степенях в обеих частях равенства, получаем

$$n^3: \quad \frac{1}{3} = \frac{1}{3},$$

$$n^2: \quad 0 = -\frac{3}{6} + \frac{1}{2},$$

$$n: \quad 0 = \frac{1}{6} - 2A_1,$$

$$1: \quad 0 = 2A_1 + 4A_2.$$

Последние два уравнения дают

$$A_1 = \frac{1}{12}, \quad A_2 = -\frac{1}{24}.$$

Подставляя $f(x) = x^4$, получим подобным образом

$$A_3 = \frac{19}{720}, \quad A_4 = -\frac{3}{160},$$

а при $f(x) = x^6$ имеем

$$A_5 = \frac{863}{60480}, \quad A_6 = -\frac{275}{24192}.$$

Итак, мы получаем формулу Грегори

$$\begin{aligned} \frac{1}{h} \int_0^{nh} f(x) dx &= \left(\frac{1}{2} f_0 + f_1 + f_2 + \dots + f_{n-1} + \frac{1}{2} f_n \right) + \\ &+ \frac{1}{12} [\Delta f_0 - \Delta f_{n-1}] - \frac{1}{24} [\Delta^2 f_0 + \Delta^2 f_{n-1}] + \\ &+ \frac{19}{720} [\Delta^3 f_0 - \Delta^3 f_{n-1}] - \frac{3}{160} (\Delta^4 f_0 + \Delta^4 f_{n-1}) + \dots \end{aligned}$$

Часто более полезна формула Грегори в форме Лагранжа, в которой линейный оператор выписан через значения функции; именно с нее мы начали этот параграф. Коэффициенты этой формы даны в

следующей таблице, где столбец слева дает коэффициент A_k наивысшего порядка. По симметрии нам нужен лишь один конец формулы:

$$A_0: \frac{1}{2} \quad 1 \quad 1 \quad 1 \quad 1 \quad \dots$$

$$A_1: \frac{5}{12} \quad \frac{13}{12} \quad 1 \quad 1 \quad 1 \quad \dots$$

$$A_2: \frac{9}{24} \quad \frac{28}{24} \quad \frac{23}{24} \quad 1 \quad 1 \quad \dots$$

$$A_3: \frac{251}{720} \quad \frac{897}{720} \quad \frac{633}{720} \quad \frac{739}{720}$$

Соответствующая формула, которая использует производные вместо разностей, известна как разложение Эйлера — Маклорена и иногда бывает полезна, но при использовании на машине она требует программирования производных (см. уравнения (4.8-2)).

Упражнения

10.10-1. Используя методы этого параграфа, вывести соответствующие формулы для

$$\int_0^n f(x) dx = af\left(\frac{1}{2}\right) + bf\left(\frac{3}{2}\right) + cf\left(\frac{5}{2}\right) + \dots + cf\left(n - \frac{5}{2}\right) + \\ + bf\left(n - \frac{3}{2}\right) + af\left(n - \frac{1}{2}\right).$$

10.10-2. Используя те же узлы, что и в упражнении 10.10-1, выполнить работу по нахождению коэффициентов первых четырех разностных членов.

§ 10.11. Выводы

Основной целью этой главы было показать, как выводить различного рода формулы; мы не слишком углублялись в рассмотрение имеющихся результатов. Одной из причин этого является обилие формул. Их существует значительно больше, чем можно поместить в одной книге, и исследование всех результатов может обрасти такой массой деталей, что обескуражит даже самого пылкого любителя численного анализа. Когда методы нахождения формул поняты, легко найти конкретную формулу, необходимую в конкретном случае.

Мы отмечаем также возможность действительно выполнять на машине большую часть работы по выведению формул. Чтобы помочь этому, мы даем блок-схему (рис. 10.11-1) метода вывода формул, не имеющих линейных ограничений на веса.

Здесь n — количество узлов x_i , n_1 — количество фиксированных x_i , $2n - n_1$ — минимальное количество моментов m_k , заданных машине (при случае машина будет требовать дальнейших моментов).

Соответствующая блок-схема для случая заданных линейных связей не включена сюда из-за неопределенности типов формул, которые

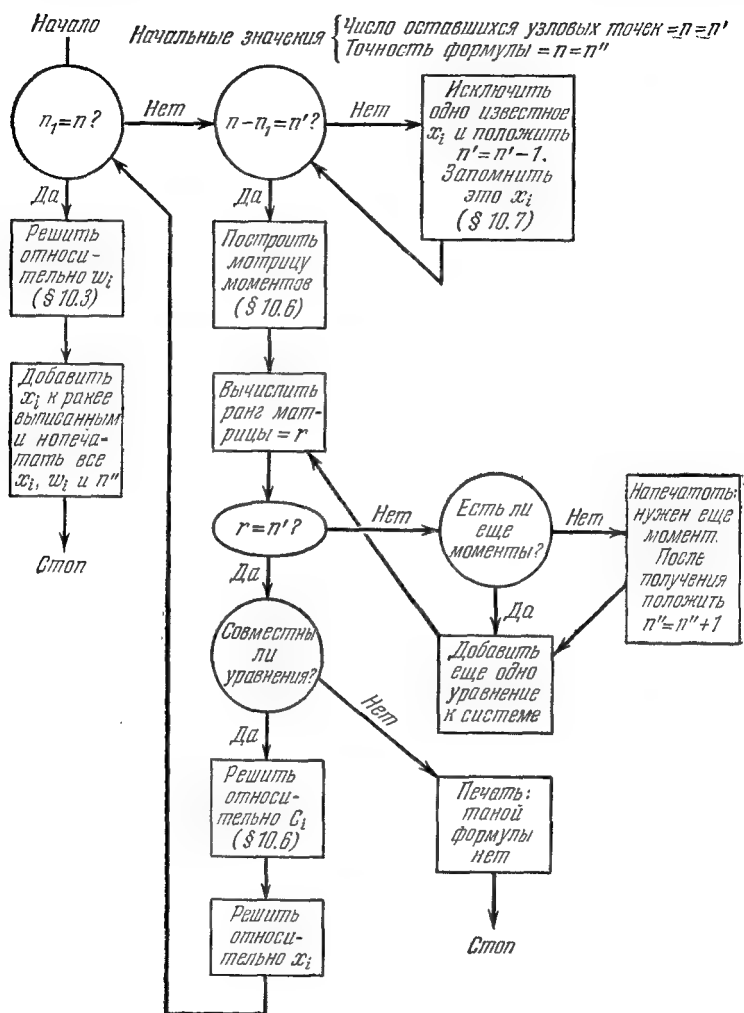


Рис. 10.11-1.

могли бы понадобиться. Как было замечено раньше, линейные связи, вероятно, должны налагаться на веса для фиксированных узлов или, как при интегрировании по Чебышеву, единообразно на все веса свободных узлов.

ГЛАВА II

О НАХОЖДЕНИИ ОСТАТОЧНОГО ЧЛЕНА ФОРМУЛЫ

§ 11.1. Потребность в остаточном члене

Описав до некоторой степени единый метод нахождения формул, рассмотрим теперь единый метод нахождения соответствующих остаточных членов. Как и в случае нахождения формул, единый метод не всегда самый короткий и элегантный, но его достоинства, по-видимому, перевешивают большую иногда трудоемкость.

В этой главе мы сосредоточим внимание на нахождении формулы остаточного члена, игнорируя эффекты округления. Как и в случае интерполяции многочленом, остаточный член будет записан в виде некоторой производной (или производных) высшего порядка. Это — традиционная форма, но она имеет все недостатки, обсуждавшиеся в § 8.7.

Здесь мы займемся выводом остаточного члена и отложим обсуждение полученных результатов до следующей главы, где собраны вместе некоторые формулы и их остаточные члены. Очевидно, читателю должна быть ясна необходимость остаточного члена. Прежде чем предпочесть одну формулу другой, мы хотели бы знать их остаточные члены, хотя величина остаточного члена и не является единственным критерием для выбора формулы.

§ 11.2. Порядок остаточного члена

В качестве примера линейного оператора будем снова использовать формулу приближенного вычисления определенного интеграла, хотя метод работает и в случае неопределенного интегрирования, интерполяции и некоторых других линейных операторов. Обозначим через $E_k = E_k(x)$ ошибку формулы, когда мы подставляем x^k вместо $f(x)$. Наш метод вывода формулы показывал (кроме формулы Грегори), что $E_k = 0$ для $k = 0, 1, \dots, m-1$, где m — количество свободных параметров в формуле. Мы ожидаем, что $E_m \neq 0$, но готовы к тому, что оно равно нулю, так как мы уже знаем, что в некоторых случаях, таких как формула Симпсона, точность выше, чем количество свободных параметров.

Проверим, равно ли E_m нулю. В том случае, когда все узловые точки заданы, E_m есть рациональная функция данных моментов и данных узловых точек. Но в случае квадратуры Гаусса, например, необходимо решить алгебраическое уравнение для узловых точек x_j . Узловой многочлен, корни которого нужно найти, чтобы получить узловые точки, дает соотношение, позволяющее выразить степени x_j , большие n (количество неизвестных узловых то-

чек), через степени, меньшие n . Это выражение включает коэффициенты C_k узлового многочлена. Теперь степени x_j , меньшие n , умножаются на их веса w_j и, следовательно, выражаются через моменты m_k . В результате опять получается рациональная функция от моментов *). Проверка $E_m = 0$ является, таким образом, рациональной операцией над данными моментами и данными узлами, и если можно выполнить достаточно много алгебраических действий с рациональными выражениями, то можно решить вопрос о том, равно или нет E_m нулю, без помех округления, которые возникают при нахождении корней узлового многочлена.

Если мы найдем, что одно значение E_m равно нулю, то попробуем следующее, и так до тех пор, пока не найдем такое m , что $E_m \neq 0$, или же мы устанем и бросим; последнее маловероятно, кроме тривиальных случаев, так как одна или две степени дополнительной точности — это обычно все, что можно получить.

Упражнения

11.2-1. Определить m для равенства (10.2-8), используя методы § 11.2.

11.2-2. Определить m для формулы Симпсона (10.4-1).

11.2-3. Определить m для упражнения 10.2-1.

§ 11.3. Функция влияния

Уточним сначала, какой класс формул мы собираемся исследовать. Для определенности будем говорить об операции интегрирования, хотя допустимы и другие линейные операторы.

В правой части равенства используются значения функции и ее производных; но мы будем требовать, чтобы старшая производная имела порядок меньший, чем точность формулы, которая в предыдущем параграфе обозначалась через $m - 1$. Таким образом, рассматриваются формулы вида

$$\int_a^b f(x) dx = \sum_{k=1}^n a_k f(x_k) + \sum_{k=1}^n b_k f'(x_k) + \dots + \sum_{k=1}^n m_k f^{(m-2)}(x_k) + R, \quad (11.3-1)$$

где некоторые из a_k, b_k, \dots, m_k могут быть нулями. Следует заметить, что узловые точки x_k могут лежать внутри или снаружи области интегрирования. Предполагается, что эта формула точна для 1, x, \dots, x^{m-1} и $E_m \neq 0$, где m определено, как в предыдущем параграфе.

*) Моменты m_k не обязаны, конечно, быть рациональными, и, вообще говоря, может возникнуть вопрос о том, уничтожаются или нет все члены в выражении E_m через моменты; но на практике это обычно не причиняет беспокойства.

Чтобы получить выражение для остаточного члена R , начнем с ряда Тейлора

$$f(x) = f(A) + (x-A)f'(A) + \frac{(x-A)^2}{2!} f''(A) + \dots + \\ + \frac{(x-A)^{m-1}}{(m-1)!} f^{(m-1)}(A) + \frac{1}{(m-1)!} \int_A^x f^{(m)}(s) (x-s)^{m-1} ds. \quad (11.3-2)$$

Эта формула может быть проверена интегрированием последнего члена по частям m раз (таким образом, она представляет собой просто тождество). Мы выбираем A как минимум из a и всех x_k . Подставим теперь выражение (11.3-2) в формулу (11.3-1) в обе части, т. е. вместо $f(x)$ и вместо $f'(x_k)$, $f''(x_k)$, ..., $f^{(m-2)}(x_k)$. Ошибка $R(f)$ в формуле (11.3-1) есть разность между двумя частями равенства. Сумма всех членов разложения Тейлора, кроме интегрального, есть многочлен от x степени не выше $m-1$. А так как формула (11.3-1) является точной для $1, x, \dots, x^{m-1}$ (т. е. $E_k = 0$; $k = 0, 1, \dots, m-1$), то после подстановки останутся лишь интегральные члены. Мы имеем, следовательно,

$$R = R(f) = \int_a^b \frac{1}{(m-1)!} \int_A^x f^{(m)}(s) (x-s)^{m-1} ds dx - \\ - \sum \frac{a_k}{(m-1)!} \int_A^{x_k} f^{(m)}(s) (x_k-s)^{m-1} ds - \\ - \sum \frac{b_k}{(m-1)!} \frac{d}{dx_k} \int_A^{x_k} f^{(m)}(s) (x_k-s)^{m-1} ds - \dots \\ \dots - \sum \frac{m_k}{(m-1)!} \frac{d^{m-2}}{dx_k^{m-2}} \int_A^{x_k} f^{(m)}(s) (x_k-s)^{m-1} ds. \quad (11.3-3)$$

Упростим (11.3-3) с помощью следующего приема. Определим для $j > 0$

$$(x-s)_+^j = \begin{cases} 0, & \text{если } x-s \leq 0, \\ (x-s)^j, & \text{если } x-s \geq 0. \end{cases} \quad (11.3-4)$$

Если $j=0$, то

$$(x-s)_+^0 = \begin{cases} 0, & \text{если } x-s < 0, \\ 1, & \text{если } x-s > 0, \end{cases}$$

но $j=0$ не может встретиться в нашем случае, так как старшая производная имеет порядок $m-2$

Легко видеть, что

$$\begin{aligned}\frac{d}{dx}(x-s)_+^{m-1} &= (m-1)(x-s)_+^{m-2}, \\ \frac{d^2}{dx^2}(x-s)_+^{m-1} &= (m-1)(m-2)(x-s)_+^{m-3}\end{aligned}\quad (11.3-5)$$

и т. д., до тех пор, пока порядок производной меньше, чем $m-1$.

Кроме того,

$$\int_a^b (x-s)_+^{m-1} dx = \frac{(b-s)_+^m - (a-s)_+^m}{m}. \quad (11.3-6)$$

Используя эти новые обозначения, мы можем увеличивать верхний предел интегрирования по s до B в уравнении (11.3-3), где B выбирается как максимум из a и всех x_k . Это увеличение в верхнем пределе не влияет на значение интеграла, так как каждый член есть нуль при $s > x_k$.

Учитывая эти замечания и выполняя дифференцирование с использованием (11.3-5) и интегрирование с использованием (11.3-6), получаем для (11.3-3)

$$\begin{aligned}R = R(f) &= \frac{1}{(m-1)!} \int_A^B f^m(s) \left[\frac{(b-s)_+^m - (a-s)_+^m}{m} - \right. \\ &- \sum_{k=1}^n a_k (x_k - s)_+^{m-1} - \sum_{k=1}^n b_k (m-1)(x_k - s)_+^{m-2} - \dots \\ &\left. \dots - \sum_{k=1}^n m_k (m-1)(m-2)\dots 3 \cdot 2 (x_k - s)_+ \right] ds. \quad (11.3-7)\end{aligned}$$

Произведение $\frac{1}{(m-1)!}$ на выражение в квадратных скобках зависит от s и исследуемой формулы, но не зависит от функции $f(s)$. Таким образом, пишем

$$R = R(f) = \int_A^B f^{(m)}(s) G(s) ds; \quad (11.3-8)$$

$G(s)$ называется «функцией влияния». В частности, $f(x) = x^m$ дает возможность вычислить интеграл от $G(s)$ из (11.3-7):

$$R(f) = R(x^m) = m! \int_A^B G(s) ds = E_m \neq 0. \quad (11.3-9)$$

Замечим, что эта функция выведена для (11.3-1) и неприменима к другим видам формул.

Исследование $G(s)$ показывает, что она определена на $[A, B]$ и является многочленом степени $\leq m$ на каждом отрезке, ограниченном соседними узловыми точками и пределами интегрирования a, b .

§ 11.4. Случай, когда $G(s)$ имеет постоянный знак

В большинстве важных случаев оказывается, что функция влияния $G(s)$ имеет постоянный знак. Если это так, то остаточный член (формула (11.3-8))

$$R = R(f) = \int_A^B f^{(m)}(s) G(s) ds$$

удовлетворяет условиям теоремы о среднем значении, и, используя (11.3-9), можно написать

$$R = f^{(m)}(\theta) \int_A^B G(s) ds = \frac{E_m}{m!} f^{(m)}(\theta) \quad (A \leq \theta \leq B).$$

Теперь встает задача — узнать, имеет ли $G(s)$ постоянный знак. Для начала рассмотрим ошибку правила трапеций

$$R(f) = \int_0^1 f(x) dx - \frac{f(1) + f(0)}{2},$$

которая является точной для $f(x) = 1$ и для x . Таким образом, $m = 2$. Тогда из (11.3-7) получаем

$$1! G(s) = \frac{(1-s)_+^2 - (-s)_+^2}{2} - \frac{(1-s)_+ + (-s)_+}{2}.$$

В этом случае $A = a = x_1 = 0$; $B = b = x_2 = 1$; $G(s)$ является многочленом на всем отрезке $[A, B]$, а именно:

$$G(s) = \frac{(1-s)^2 - (1-s)}{2} = \frac{-s + s^2}{2} = -\frac{s(1-s)}{2} < 0,$$

который, очевидно, имеет постоянный знак (рис. 11.4-1).

В качестве второго примера рассмотрим формулу Симпсона

$$\int_{-1}^1 f(x) dx = \frac{f(-1) + 4f(0) + f(1)}{3}.$$

Краткое исследование показывает, что $m = 4$; следовательно, используя (11.3-7),

$$3! G(s) = \frac{(1-s)_+^4 - (-1-s)_+^4}{4} - \frac{(-1-s)_+^3 + 4(-s)_+^3 + (1-s)_+^3}{3}.$$

Здесь $A=a=x_1=-1$; $B=b=x_3=1$; $x_2=0$. Следовательно, нужно рассматривать два отрезка: $[-1, 0]$ и $[0, +1]$. $1 > s > 0$:

$$3! G(s) = \frac{(1-s)^4}{4} - \frac{(1-s)^3}{3} = (1-s)^3 \left(\frac{-1-3s}{12} \right) < 0,$$

$$0 > s > -1:$$

$$\begin{aligned} 3! G(s) &= \frac{(1-s)^4}{4} - \frac{(1-s)^3}{3} - \frac{4(-s)^3}{3} = \frac{-1+6s^2+8s^3+3s^4}{12} = \\ &= (1+s)^3 \left(\frac{-1+3s}{12} \right) < 0 \end{aligned}$$

(рис. 11.4-2).

Эти примеры показывают, что если формула симметрична, то $G(s)$ также симметрична. Доказательство вытекает из замечания, что если

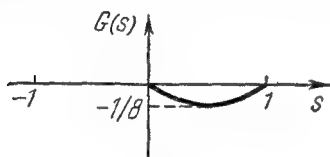


Рис. 11.4-1.

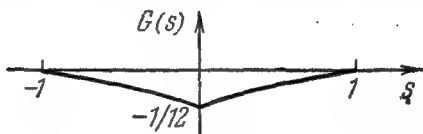


Рис. 11.4-2.

формула симметрична, то мы должны получить ту же самую ошибку, проинтегрировав $f(x)$ и $f(-x)$. А так как мы имеем свободный выбор функции $f(x)$, любое отсутствие симметрии в $G(x)$ может быть использовано, чтобы получить разные ответы.

Функция влияния $G(s)$ определена на отрезке $[A, B]$, который разбит на интервалы узлами x_k и пределами интегрирования a и b . Она непрерывна на $[A, B]$ и на каждом подынтервале является многочленом степени $\leq m$ (для однократных интегралов; для повторного интегрирования эта степень может быть и выше).

Задача сводится к тому, чтобы определить, имеют ли эти многочлены нули в своих подынтервалах. Прямой подход состоит в том, чтобы найти нули этих многочленов и затем проверить, лежат ли они в соответствующих подынтервалах. Однако этот способ может потребовать много машинного времени. С другой стороны, для больших промежутков вычисление $G(s)$ в близких точках и вычерчивание графика едва ли возможны, хотя простота этого метода очевидна. Алгебраический метод, который был применен в приведенных примерах, вероятно, труден для программирования на машине, хотя для ручной проверки он часто пригоден. Таким образом, вполне удовлетворительный способ проверки, меняет ли знак $G(s)$, вряд ли существует. Но, во всяком случае, следует ясно представлять себе, что $G(s)$ зависит лишь от формулы и, следовательно, проверка должна проводиться для любой данной формулы лишь один раз.

Упражнения

11.4-1. Проверить, что формула в упражнении (10.2-1) для интеграла

$$\int_{-\pi}^{\pi} f(x) \sin x dx$$

имеет функцию влияния $G(s)$ постоянного знака, хотя ядро $K(x) = \sin x$ меняет знак.

11.4-2. Проверить формулу упражнения (10.2-2).

§ 11.5. Случай, когда функция влияния меняет знак

Если $G(s)$ меняет знак, то следует отметить три обстоятельства. Во-первых, тогда существуют функции $f(x)$, для которых нет такого θ , чтобы была верной теорема о среднем значении. Чтобы показать это, предположим, что $G(s)$ является положительной всюду, кроме маленького интервала, в котором она отрицательна, и что интеграл от $G(s)$ положителен. Рассмотрим такую функцию $f(x)$, что $f^{(m)}(x)$ положительна и непрерывна, но мала вне этого маленького интервала и очень велика внутри. Тогда интеграл

$$\int_A^B f^{(m)}(s) G(s) ds \quad (11.5-1)$$

отрицателен, но $f^{(m)}(\theta)$ и интеграл от $G(s)$ оба положительны. Следовательно, не существует нужного θ . Подобные аргументы могут быть приведены для других $G(s)$.

Во-вторых, в этом случае можно оценить сверху модуль интеграла (11.5-1). Для $|G(s)|$ верна теорема о среднем, откуда легко получается оценка:

$$\begin{aligned} \left| \int_A^B f^{(m)}(s) G(s) ds \right| &\leq \int_A^B |f^{(m)}(s)| |G(s)| ds = \\ &= |f^{(m)}(\theta)| \cdot \int_A^B |G(s)| ds \leq \max |f^{(m)}(s)| \cdot \int_A^B |G(s)| ds, \end{aligned}$$

где максимум взят по всем x ($A \leq x \leq B$). Интеграл

$$\int_A^B |G(s)| ds$$

можно вычислить аналитически или численно. Это опять необходимо сделать только один раз для любой данной формулы.

В-третьих, хотя знак $G(s)$ меняется, мы все же можем найти остаточный член. В качестве иллюстрации рассмотрим функцию $G(s)$

на рис. 11.5-1. Точка $x=C$ выбрана так, что две заштрихованные площади равны по величине, т. е.

$$\int_C^B G(s) ds = 0.$$

Мы имеем (см. уравнения (11.3-8) или (11.5-1))

$$R = R(f) = \int_A^B f^{(m)}(s) G(s) ds = \int_A^C + \int_C^B.$$

Проинтегрируем второй член по частям и положим $\int_C^s G(s) ds = H(s)$.
(Замечание: $H(B) = H(C) = 0$.)

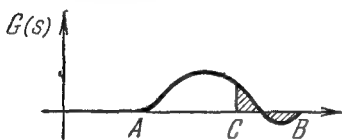


Рис. 11.5-1. Частный вид $G(s)$.

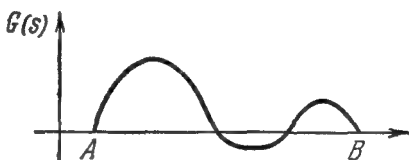


Рис. 11.5-2. Другой вид $G(s)$.

Тогда

$$\begin{aligned} R &= \int_A^C f^{(m)}(s) G(s) ds + f^{(m)}(s) H(s) \Big|_C^B - \int_C^B f^{(m+1)}(s) H(s) ds = \\ &= \int_A^C f^{(m)}(s) G(s) ds - \int_C^B f^{(m+1)}(s) H(s) ds. \end{aligned}$$

Теперь можно применить теорему о среднем значении к каждому интегралу

$$R = f^{(m)}(\theta_1) \int_A^C G(s) ds - f^{(m+1)}(\theta_2) \int_C^B H(s) ds.$$

Если $G(s)$ такая, как показано на рис. 11.5-2, то можно просто положить

$$H(s) = \int_A^s G(s) ds$$

и интегрировать (11.5-1) по частям

$$\begin{aligned} R(f) &= f^{(m)}(s) H(s) \Big|_A^B - \int_A^B f^{(m+1)}(s) H(s) ds = \\ &= f^{(m)}(B) H(B) - f^{(m+1)}(\theta) \int_A^B H(s) ds. \end{aligned}$$

Мы не будем анализировать общий случай, так как эти результаты не используются в тексте. Заметим только, что такие формы остаточного члена могут быть иногда полезны.

§ 11.6. Слабое место в методе рядов Тейлора

Так как существует остаточный член для рядов Тейлора в виде *)

$$f(a+x) = f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2!}f''(a) + \dots \\ \dots + \frac{(x-a)^{m-1}}{(m-1)!}f^{(m-1)}(a) + \frac{(x-a)^m f^{(m)}(a+\theta x)}{m!},$$

естественно попытаться использовать метод рядов Тейлора (второй метод § 7.2), чтобы получить ошибку формулы.

Пусть используется метод рядов Тейлора. Коэффициенты формулы определены так, чтобы степени x уничтожались с обеих сторон вплоть до степени m , которая по предположению не уничтожается. Зафиксируем мысленно значение x . Тогда получим ряд из членов $f^{(m)}(a+\theta_i x_i)$ для различных θ_i

$$\alpha_1 f^{(m)}(a+\theta_1 x_1) + \alpha_2 f^{(m)}(a+\theta_2 x_2) + \dots + \alpha_k f^{(m)}(a+\theta_k x_k),$$

где, конечно, θ_i зависят от x_i , $\theta_i = \theta_i(x_i)$. Мы хотим теперь заменить это выражение таким:

$$(\alpha_1 + \alpha_2 + \dots + \alpha_k) f^{(m)}(a+\theta x)$$

для каких-то новых значений θ (и ожидаем, что многие члены α_i уничтожатся). При каких обстоятельствах существуют такие θ ?

Как известно из метода функции влияния, если $G(s)$ сохраняет знак, то такое θ существует, а если у $G(s)$ знак не постоянный, то существуют некоторые функции $f(x)$, для которых такого θ нет. Таким косвенным путем можно ответить на вопрос. Кажется, прямого способа найти ответ нет, но если ответ найден, то есть возможность действовать двумя различными способами, а такой выбор иногда полезен.

В случае формулы гауссова вида с несколькими свободными узловыми точками существует специальная техника **) определения остаточного члена проще, чем через функцию $G(x)$, и это наводит на мысль, что есть некоторая надежда найти другой единый метод.

*) Это выражение называют формой Лагранжа для остаточного члена, а иногда все выражение называется расширенной теоремой о среднем.

**) См. Хильдебранд [14] или Копал [20].

ГЛАВА 12

ФОРМУЛЫ ДЛЯ ОПРЕДЕЛЕННЫХ ИНТЕГРАЛОВ

§ 12.1. Введение

Между вычислением определенного интеграла

$$\int_a^b f(x) dx$$

и неопределенного интеграла, который можно записать в виде

$$\int_a^x f(x) dx,$$

имеется существенное различие. Результатом вычисления первого является одно число, тогда как результатом второго является таблица. В этой главе будет рассмотрено вычисление определенных интегралов. Следующая глава посвящена неопределенным.

Определенный интеграл можно вычислять, как применяя единую формулу на всем отрезке, так и разбивая отрезок интегрирования и применяя формулу к каждой из его частей. Примером последнего является широко используемая формула Симпсона

$$\int_a^{a+2nh} f(x) dx = \frac{h}{3} \left[f(a) + 4f(a+h) + 2f(a+2h) + 4f(a+3h) + \dots \right. \\ \left. \dots + f(a+2nh) \right], \quad (12.1-1)$$

которая выводится последовательным применением формулы

$$\int_a^{a+2h} f(x) dx = \frac{h}{3} \left[f(a) + 4f(a+h) + f(a+2h) \right]$$

к каждому двойному интервалу.

Обычное соображение в пользу применения единой формулы для всего отрезка заключается в том, что для одинакового числа узловых точек такая формула имеет остаточный член более высокого порядка. Но если в комплексной плоскости вблизи интервала интегрирования есть особенности подынтегральной функции, то остаточный член высокого порядка скорее всего будет большим. Конечно, при $h \rightarrow 0$ остаточный член более точной формулы стремится к нулю быстрее, чем для формулы низкой точности. К сожалению, место, где это преимущество начинает работать, обычно неизвестно, даже если интеграл вычислен при двух различных разбиениях, когда, например, второе

включает первое. Поэтому на практике часто предпочитают составную формулу.

Другим основанием для предпочтения одной формулы другой является влияние округления узловых значений подынтегральной функции на результат интегрирования. Это влияние измеряется суммой квадратов весов. Если формула должна быть точной, когда $f(x)$ константа, то сумма весов w_i постоянна. Требуется минимизировать

$\sum_{i=1}^n w_i^2$ при условии, что сумма w_i постоянна. Как известно, лучше

всего взять все w_i равными между собой. Любое отклонение от равенства вызывает некоторое увеличение ошибок округления сверх минимума. Однако этот эффект не так велик, как можно было бы предполагать. Например, в составной формуле Симпсона (12.1-1) колебания в весах вдвое дают среднеквадратичный эффект

$$\frac{(4/3)^2 + (2/3)^2}{2} = \frac{10}{9} = 1,111$$

по сравнению с единицей, которая получается, когда все веса равны. Если некоторые веса отрицательны, то суммарная погрешность может быть значительно больше, как, например, в формулах Ньютона — Котеса (см. § 12.2) высокого порядка.

Задача численного интегрирования, как было замечено в § 7.3, есть задача выбора узловых точек, по которым можно оценить среднее значение подынтегральной функции. Изложенные до сих пор в этой книге методы состояли в аппроксимации многочленом подынтегральной функции, или множителя подынтегральной функции, и последующем интегрировании многочлена. Таким образом, подынтегральная функция $f(x)$ рассматривается как произведение двух сомножителей: $K(x)$ и $g(x)$. При этом мы требуем выполнения трех свойств. Во-первых, функция $K(x)$ не должна изменяться за все время пользования формулой; она может, конечно, быть просто единицей. Во-вторых, моменты m_k

$$m_k = \int_a^b K(x) x^k dx \quad (12.1-2)$$

должны находиться аналитически. В-третьих, функция $g(x)$ должна допускать достаточно хорошую аппроксимацию многочленом степени n . При выполнении этих требований машина может вывести формулу для вычисления интеграла. Методы гл. 10 показывают, что искомая формула существует и может быть найдена машиной, если задать узлы, при условии, что любая производная в узловой точке задается вместе со всеми производными низших порядков. Если некоторые или все узлы не заданы и выбираются методом интегрирования, то существует риск, что какие-либо узловые точки могут оказаться комплекс-

ными числами. Это, однако, не должно быть непреодолимым препятствием к применению формулы, хотя, конечно, несколько обескураживает.

В качестве примера применения вышеупомянутого метода рассмотрим вычисление интеграла

$$I = - \int_0^1 \frac{\ln x}{1+e^x} dx. \quad (12.1-3)$$

Множитель $\ln x$ имеет особенность при $x=0$, и, следовательно, вся подынтегральная функция не приближается многочленом в этом интервале. Положив $K(x) = \ln x$, получим моменты

$$m_k = - \int_0^1 x^k \ln x dx = \frac{1}{(k+1)^2}. \quad (12.1-4)$$

После этого остается аппроксимировать многочленом функцию $\frac{1}{1+e^x}$, что сделать нетрудно, просто взяв интерполяционный многочлен, как это делалось во второй части, или разложив эту функцию в ряд Маклорена и проинтегрировав почленно.

К сожалению, нельзя в явном виде сформулировать условия, при выполнении которых функция может быть легко аппроксимирована многочленом. Остаточные члены на практике обычно трудно найти, и нам придется решать вопрос о том, как разбить $f(x)$ на два множителя $K(x)$ и $g(x)$, из других соображений.

Прежде всего, если $f(x)$ имеет особенность, то, очевидно, она должна войти в $K(x)$. Другое соображение состоит в том, что многочлен степени n не может иметь больше чем n корней и поэтому не может достаточно хорошо приближать сильно колеблющуюся функцию. И наконец, следует учесть, что многочлен не остается ограниченным в окрестности бесконечности. До некоторой степени эти рассуждения вместе с условиями, что $K(x)$ не должна меняться от раза к разу и что моменты $K(x)$ нужно уметь находить аналитически, определяют выбор $K(x)$. Изложив эти общие соображения, обратимся к рассмотрению некоторых конкретных формул. Перечисление конкретных формул отчасти дает основу для вывода новых формул. С другой стороны, их обычно хватает для решения относительно простых задач.

Упражнения

12.1-1. Используя универсальные матрицы § 10.3, найти пятиточечную формулу для

$$- \int_0^1 f(x) \ln x dx.$$

(З а м е ч а н и е: привести область интегрирования к интервалу $(-2, 2)$ и затем вычислить моменты, используя (12.1-4).)

12.1-2. Найти коэффициенты для

$$\int_{-1}^1 \frac{f(x)}{\sqrt{1-x^2}} dx = a_{-1}f_{-1} + a_{-1/2}f_{-1/2} + a_0f_0 + a_{1/2}f_{1/2} + a_1f_1.$$

12.1-3. Рассмотреть формулу упражнения 10.2-3 как составную. Сравнить с формулой Симпсона.

12.1-4. Доказать, что если $\sum w_i = C$, то $\sum w_i^2 = \min$ тогда и только тогда, когда все w_i равны между собой.

§ 12.2. Формулы Ньютона — Котеса

Логически самые простые формулы для интегрирования — те, которые используют множество равноотстоящих узлов, — получаются интегрированием многочлена, проходящего через узлы. Узлы *) 0, h , $2h$, ..., nh приводят к формулам Ньютона — Котеса

$$\int_0^{nh} f(x) dx = w_0 f(0) + w_1 f(h) + w_2 f(2h) + \dots + w_n f(nh).$$

Первый частный случай формулы Ньютона — Котеса — это хорошо известное правило трапеций

$$\int_0^h f(x) dx = \frac{h}{2} [f(0) + f(h)] - \frac{1}{12} h^3 f''(\theta), \quad (12.2-1)$$

второй — формула Симпсона:

$$\int_0^{2h} f(x) dx = \frac{h}{3} [f(0) + 4f(h) + f(2h)] - \frac{1}{90} h^5 f^{IV}(\theta); \quad (12.2-2)$$

третий — правило трех восьмых:

$$\int_0^{3h} f(x) dx = \frac{3h}{8} [f(0) + 3f(h) + 3f(2h) + f(3h)] - \frac{3}{80} h^5 f^{IV}(\theta). \quad (12.2-3)$$

Таблица 12.2-1 дает коэффициенты для формулы

$$\int_0^{nh} f(x) dx = Ah [B_0 f(0) + B_1 f(h) + B_2 f(2h) + \dots + B_n f(nh)] + K_n$$

где n — число интервалов, а не число узловых точек. Коэффициенты симметричны, и, следовательно, достаточно дать лишь часть таблицы.

*) Заметим, что здесь используется $n+1$ узловая точка, а не n , как это делалось раньше.

Таблица 12.2-1

n	A	B_0	B_1	B_2	B_3	B_4	B_5	K_n
1	1/2	1	1					$-(1/12) h^3 f''(\theta)$
2	1/3	1	4	1				$-(1/90) h^5 f^{IV}(\theta)$
3	3/8	1	3	3	1			$-(3/80) h^5 f^{IV}(\theta)$
4	2/45	7	32	12	32	7		$-(8/945) h^7 f^{VI}(\theta)$
5	5/288	19	75	50	50	75	19	$-(275/12096) h^7 f^{VI}(\theta)$
6	1/140	41	216	27	272	27	216	$-(9/1400) h^9 f^{VIII}(\theta)$
7	7/17 280	751	3577	1323	2989	2989	1323	$-(8183/518400) h^9 f^{VIII}(\theta)$
8	4/14 175	989	5888	—928	10496	—4540	10496	
9	9/89 600	2857	15741	1080	19344	5778	5778	
10	5/299 376	16067	106300	—48525	272400	—260550	427368	

Некоторые коэффициенты становятся отрицательными при $n=8$. Для $n=9$ они все положительны, но для $n \geq 10$ существуют и отрицательные коэффициенты. Это приводит к росту ошибок округления; по этим причинам формулы Ньютона — Котеса более высокого порядка используются редко. Порядок остаточных членов возрастает на два при переходе от нечетного номера к следующему четному, так что формулы четного порядка предпочтительнее.

Обратим внимание на появление степеней h в остаточном члене. Если брать не единичное разбиение, то степени h автоматически возникнут в определяющих уравнениях (10.3-1). Можно либо провести выкладки с настоящим h , либо считать, что $h=1$, и соответствующие степени h вписать потом (для равноотстоящих узлов); обычно лучше последнее.

Если надо сравнить эти формулы для одного и того же интервала, то следует помнить, что h равно длине интервала интегрирования, деленной на n , и внести соответствующие поправки в остаточные члены.

Формулы Ньютона — Котеса могут быть получены многими способами. Вероятно, самый простой путь нахождения коэффициентов основан на том обстоятельстве, что формула Грегори, написанная с включением всех разностей, которые могут быть вычислены по узлам, является точной для многочленов максимальной степени и, следовательно, в форме Лагранжа она должна быть такой же, как если бы она была выведена непосредственно. Таким образом, для случая $n=4$ имеем

$$\begin{aligned} \int_0^4 f(x) dx &= \frac{1}{2} f_0 + f_1 + f_2 + f_3 + \frac{1}{2} f_4 + \frac{1}{12} (\Delta f_0 - \Delta f_3) - \\ &- \frac{1}{24} (\Delta^2 f_0 + \Delta^2 f_2) + \frac{19}{720} (\Delta^3 f_0 - \Delta^3 f_1) - \frac{3}{160} (\Delta^4 f_0 + \Delta^4 f_1) = \\ &= \frac{14}{45} f_0 + \frac{64}{45} f_1 + \frac{24}{45} f_2 + \frac{64}{45} f_3 + \frac{14}{45} f_4. \end{aligned}$$

Упражнения

12.2-1. Применить формулы Ньютона — Котеса для $n=2, 4, 6, 8$ к $\int_0^1 e^{-x} dx$. Сравнить с правильным ответом.

12.2-2. Рассмотреть рост погрешности округления в формулах Ньютона — Котеса.

12.2-3. Вывести формулу Ньютона — Котеса из формулы Грегори при $n=6$.

§ 12.3. Использование формулы Грегори

Как было показано, отбирая нужным образом члены в формуле Грегори, можно получить коэффициенты формулы Ньютона — Котеса. Это объясняет, почему формула Грегори так полезна на практике;

она и гибкая и точная. Часто, чтобы избежать отрицательных коэффициентов (которые дают большой рост погрешности округления) и производных высокого порядка в остаточном члене (который может быть очень велик), применяют составные формулы, как, например, составную формулу Симпсона

$$\int_0^{2nh} f(x) dx = \frac{h}{3} (f_0 + 4f_1 + 2f_2 + 4f_3 + \dots + 4f_{2n-1} + f_{2n}).$$

Если в формуле Грегори ограничиться вторыми разностями, то остаточный член будет того же порядка, а ошибки округления меньше, так как большинство коэффициентов будут близки по величине и сумма квадратов будет ближе к минимуму.

Т а б л и ц а 12.3-1

Вычисление интеграла $\int_0^{\infty} e^{-x^2} dx$

h	I
0,5	0,88622 69524 5 С точностью до 11-го десятичного знака
0,6	69254 8
0,7	69285
0,8	0,88622 72808
0,9	23 598
1,0	32 0
1,1	0,88674

Таким образом, широко используемая формула Грегори

$$\begin{aligned} \int_0^{nh} f(x) dx = & \frac{1}{2} f_0 + f_1 + \dots + f_{n-1} + \frac{1}{2} f_n + \frac{1}{12} (\Delta f_0 - \Delta f_{n-1}) - \\ & - \frac{1}{24} (\Delta^2 f_0 + \Delta^2 f_{n-2}) + \frac{19}{720} (\Delta^3 f_0 - \Delta^3 f_{n-3}) - \frac{3}{160} (\Delta^4 f_0 + \Delta^4 f_{n-4}) + \dots \end{aligned}$$

для равноудаленных узлов является наиболее полезной и гибкой. При вычислении такого интеграла, как

$$\int_{-\infty}^{\infty} e^{-x^2} dx,$$

концевые поправки к сумме равноотстоящих равновзвешенных значений функции (разности, стоящие на обоих концах формулы Грегори)

не дают никакого вклада. Этот результат установлен как теоретически, так и экспериментально при условии, что расстояние между узлами не слишком велико. Хартри*) дал таблицу результатов (см. таблицу 12.3-1), получающихся при вычислении по формуле

$$I = \int_0^{\infty} e^{-x^2} dx = h \left(\frac{1}{2} + \sum_{k=1}^{\infty} e^{-h^2 k^2} \right).$$

При $h = \frac{1}{2}$ совпадение прекрасное. Но, к сожалению, неизвестно, как с увеличением h падает точность. Анализ остаточного члена формулы Грегори — слишком трудная тема.

Упражнение 12.3-1. Написать блок-схему для формулы Грегори с автоматическим ограничением порядка разностей. Объясните условия этих ограничений.

§ 12.4. Открытые формулы

Формулы Ньютона — Котеса иногда называют замкнутыми, потому что они включают конечные точки интервала, в противоположность открытым формулам, которые не используют конечные значения:

$$\int_0^{nh} f(x) dx = w_1 f(h) + w_2 f(2h) + \dots + w_{n-1} f[(n-1)h].$$

Перепишем формулу в виде

$$\int_0^{nh} f(x) dx = Ah [B_1 f(n) + B_2 f(2h) + \dots + B_{n-1} f[(n-1)h]] + K_n.$$

Коэффициенты даны в таблице 12.4-1.

Таблица 12.4-1

n	A	B_1	B_2	B_3	B_4	K_n
3	3/2	1	1			$(3h^3/4) f''(\theta)$
4	4/3	2	-1	2		$(14/45) h^5 f^{IV}(\theta)$
5	5/24	11	1	1	11	$(95/144) h^5 f^{IV}(\theta)$
6	3/10	11	-14	26	-14	$(41/140) h^7 f^{VI}(\theta)$

Эти формулы иногда используются при численном интегрировании дифференциальных уравнений. Потребителю следует быть осторож-

*) См. также R. A. Fisher, Phil. Trans. Roy. Soc., London, A 222, 1922.

ным при использовании формул открытого типа, если вычислить значение подынтегральной функции в конце интервала не удастся, так как трудности вычисления, быть может, означают, что там имеется особенность, а тогда допущение о возможности аппроксимации подынтегральной функции многочленом может оказаться неправильным.

Между формулами открытого и замкнутого типа существуют промежуточные формулы (принадлежащие Маклорену), которые используют узлы в средних точках n интервалов

$$\int_0^{nh} f(x) dx = w_1 f\left(\frac{h}{2}\right) + w_2 f\left(\frac{3h}{2}\right) + \dots + w_n f\left(\frac{2h-1}{2} h\right).$$

Упражнение 12.4-1. Вывести формулы типа Маклорена в форме Лагранжа из упражнения 10.10-1.

§ 12.5. Квадратура Гаусса

Название «гауссова квадратура» ассоциируется со случаем, когда все узлы свободны. Существуют три широко используемых случая: отрезок $[-1, 1]$ с подынтегральным весовым множителем, равным единице (Гаусс — Лежандр); область $[0, \infty)$ с подынтегральным весовым множителем e^{-x} (Гаусс — Лагерр) и область $(-\infty, \infty)$ с подынтегральным весовым множителем e^{-x^2} (Гаусс — Эрмит). В каждом случае второе имя присоединяется потому, что узловые точки оказываются нулями соответствующих ортогональных многочленов степени n . Случаи Лежандра (n от 1 до 16) и Лагерра (n от 1 до 15) затабулированы Национальным бюро стандартов [31]*).

Мы показали, что узловые точки все различны, действительны и лежат в интервале интегрирования, а веса все положительны. Остаточный член пропорционален производной порядка $2n$

$$E = \frac{f^{(2n)}(\theta)}{(2n)!} \int_a^b \pi^2(x) dt,$$

где $\pi(x) = (x - x_1)(x - x_2) \dots (x - x_n)$.

Исследовано много других частных случаев, в которых ядро $K(x)$ имеет особенности на одном или на обоих концах области интегрирования. Особый интерес представляет случай Гаусса — Чебышева

$$\int_{-1}^1 \frac{f(x)}{\sqrt{1-x^2}} dx = \sum_{i=1}^n w_i f(x_i).$$

*) Аналогичные таблицы и для больших значений n имеются в книгах: В. И. Крылов и Л. Т. Шульгина, Справочная книга по численному интегрированию, «Наука», М., 1966 и А. С. Кронрод, Узлы и веса квадратурных формул, «Наука», М., 1964. (Прим. ред.)

Узловые точки оказываются такими:

$$x_j = \cos \frac{2j-1}{2n} \pi,$$

тогда как все веса оказываются равными

$$w_i = \frac{\pi}{n}.$$

Другие частные случаи, вероятно, лучше всего посмотреть в [20] и [28].

Гауссовы методы интегрирования весьма эффективны при приближенном вычислении интеграла по немногим узловым точкам при условии, что функция (исключая ядро) может быть хорошо аппроксимирована многочленом. Это справедливо для большинства примеров из учебников; многое зависит от близости особенностей к области интегрирования.

Широко распространено мнение, что так как методы квадратур Гаусса столь хороши, они чувствительны к незначительным ошибкам. Небольшое рассуждение показывает, что если узел задан с ошибкой, то приближенное значение интеграла изменяется на величину этой ошибки, умноженную на производную подынтегральной функции и вес в этом узле. То же самое, конечно, верно для любого метода интегрирования. Тот факт, что все веса гауссовой квадратуры положительны, обеспечивает для нее отсутствие интерференции.

Упражнение 12.5-1. Показать, что формула

$$\int_{-1}^1 \frac{f(x)}{\sqrt{1-x^2}} dx = \frac{\pi}{3} \left[f\left(-\frac{\sqrt{3}}{2}\right) + f(0) + f\left(\frac{\sqrt{3}}{2}\right) \right]$$

точна для многочленов пятой степени.

§ 12.6. Формулы интегрирования смешанного гауссового типа

Эффективность гауссового подхода к интегрированию и используемые этим методом математические соображения позволяют решить многие конкретные задачи. В настоящем параграфе будут перечислены некоторые из них. Рад исследовал случай, когда некоторые узлы заранее фиксированы *). Основной интерес представляет случай, когда в число узлов включены обе концевые точки. Причина этого интереса состоит в том, что, когда составная формула применяется ко всему отрезку, значения функции в общих концевых точках

*) Аналогичный случай рассматривался также А. С. Кронродом и за-
табулирован им для n от 1 до 40 в книге, упомянутой на стр. 169.
(Прим. ред.)

должны вычисляться лишь однажды. Копал [20] затабулировал этот случай для всех значений n от 3 до 11.

Ральстон (см. § 10.9) исследовал интересный случай, когда веса на концах отрезка отличаются только знаками. В составной формуле веса взаимно уничтожаются, за исключением двух крайних концов. Таким образом, точность улучшается благодаря свободному весовому параметру, а цена вычисления повышается лишь слегка, и чем длиннее область интегрирования, тем эта дополнительная цена относительно меньше.

Во всех этих случаях остаточный член обычно имеет вид

$$\frac{f^{(m)}(\theta)}{m!} \int_a^b k(x) \pi^2(x) Q' dx,$$

где Q' — произведение множителей в фиксированных узловых точках.

Упражнение 12.6-1. Рассмотреть применение формулы в упражнении 10.2-2 как составной формулы. Сравнить ее с составной формулой Симпсона.

§ 12.7. Суммирование рядов

Формула Грегори выражает интеграл как сумму значений подынтегральной функции, вычисленных в равноудаленных точках, с равными весами, плюс поправки, затрагивающие лишь концевые точки. Если область интегрирования уходит в бесконечность, то сумма становится бесконечным рядом.

Можно, наоборот, выразить сумму бесконечного ряда как несобственный интеграл, плюс поправочные члены в начале (такие же поправки на бесконечности обращаются в нуль вследствие сходимости ряда). Например,

$$\begin{aligned} \sum_{x=1}^{\infty} \frac{1}{x^2} &= \int_1^{\infty} \frac{dx}{x^2} + \frac{1}{2} \left(\frac{1}{1^2} \right) - \frac{1}{12} \left(\frac{1}{2^2} - \frac{1}{1^2} \right) + \frac{1}{24} \left(\frac{1}{3^2} - \frac{2}{2^2} + \frac{1}{1^2} \right) \dots = \\ &= 1 + \frac{1}{2} + \frac{3}{48} + \frac{11}{432} \dots = 1,588 \dots \text{ (верный ответ } = 1,64 \dots \text{)}. \end{aligned}$$

Часто бывает выгодно сложить несколько первых (скажем, 10) членов отдельно и применить интеграл к остальным.

Может случиться, что интеграл, возникший из ряда, не может быть вычислен аналитически. Тогда можно или проинтегрировать его численно, или сделать замену переменных и приближенно вычислить интеграл по формуле Грегори. При переходе от рядов к интегралу и обратно дважды возникают поправочные члены на концах. Выигрыш, который можно получить при таких преобразованиях, состоит в том, что получающийся новый ряд сходится гораздо быстрее первоначального.

§ 12.8. Эффекты замены переменной

В двух формулах Гаусса

$$\int_{-\infty}^{\infty} e^{-x^2} f(x) dx, \quad \int_0^{\infty} e^{-x} f(x) dx$$

изменение масштаба является единственной степенью свободы. Используя второй интеграл в качестве примера, запишем его как

$$I = \int_0^{\infty} e^{-\alpha x} [e^{-(1-\alpha)x} f(x)] dx.$$

Если положить $\alpha x = \theta$ ($\alpha > 0$),

$$I = \frac{1}{\alpha} \int_0^{\infty} e^{-\theta} F(\theta) d\theta,$$

где

$$F(\theta) = e^{-[(1-\alpha)/\alpha]\theta} f\left(\frac{\theta}{\alpha}\right),$$

то можно применить ту же формулу интегрирования к этому новому интегралу, как и к первоначальному. Выбор значения α зависит от того, насколько хорошо соответствующая $F(\theta)$ аппроксимируется многочленом заданной степени. Вообще говоря, если большинство слагаемых $\varpi_i F(x_i)$ оказались малыми, то следует думать, что α выбрано неудачно.

Такое изменение масштаба ставит вопрос о замене переменной. Снова используя

$$\int_0^{\infty} e^{-x} f(x) dx,$$

получим $x = -\ln \theta$. Тогда придем к интегралу

$$\int_0^1 f(-\ln \theta) d\theta.$$

Возникает вопрос: «какую функцию $f(x)$ или $f(-\ln \theta)$, соответственно в областях аппроксимации ($0 \leq x \leq \infty$) или ($0 \leq \theta \leq 1$) легче приблизить многочленом?»

Заключительное замечание на тему о замене переменных: обычно заменой переменных можно исключить особенности. В приведенном ранее примере (§ 1.9) рассматривался интеграл

$$g(y) = \frac{d}{dy} \int_0^y \frac{f(x)}{\sqrt{y-x}} dx.$$

Подстановка $x = y \sin^2 \theta$ приводит его к виду

$$g = \frac{d}{dy} \int_0^{\pi/2} 2 \sqrt{y} f(y \sin^2 \theta) \sin \theta d\theta,$$

т. е. устраняет особенность.

§ 12.9. Интегралы с параметром

Часто бывает необходимо вычислить интеграл вида

$$F(\lambda) = \int_0^{\infty} f(x, \lambda) dx$$

для некоторых значений параметра λ .

Опыт работы показывает, что обычно при дифференцировании по параметру можно найти дифференциальное уравнение для $F(\lambda)$, и это уравнение может быть решено в том смысле, что оно сводится к интегралу вида

$$F(\lambda) = A(\lambda) + \int_0^{\lambda} g(\lambda) d\lambda.$$

В этом виде каждый шаг вперед по λ дает другое решение $F(\lambda)$. В большинстве случаев этот метод является более эффективным подходом к задаче.

Другое соображение состоит в том, что часто интеграл от сильно осциллирующей подынтегральной функции может быть превращен в другой с быстро убывающей неколеблющейся подынтегральной функцией.

Можно сказать, что эти несколько замечаний сделаны на тему, как не вычислять интеграл. В действительности практика осуществления этих советов относится к той области классического «формульно-аналитического» математического анализа, которой редко владеют в наши дни и которая не относится к предмету нашей книги.

ГЛАВА 13

НЕОПРЕДЕЛЕННЫЕ ИНТЕГРАЛЫ

§ 13.1. Описание содержания главы и система обозначений

Основной целью этой главы является введение двух новых понятий. Первое — это неустойчивость, возникающая иногда при использовании уже вычисленных значений функции для получения ее следующего значения. Этот эффект можно сравнить с неустойчивостью в системах с «обратной связью».

Обратная связь есть использование части выхода в некоторый момент в качестве входа в несколько более позднее время. Иногда этой задержкой во времени можно пренебречь, а иногда нельзя. Вероятно, простейшим физическим примером является обратная связь усилителя (рис. 13.1-1). Выход есть сумма двух входов слева, умноженная на коэффициент усиления, взятый равным -10^9 . Мы имеем

$$\left(y + \frac{1}{10}x\right)(-10^9) = x$$

или

$$x = -\frac{10^{10}y}{10^9 + 10} = -\frac{10y}{1 + 10^{-8}} \approx -10y.$$

Таким образом, выходной сигнал равен входному, умноженному на -10 , и совершенно не зависит от небольших изменений в характеристиках усилителя (т. е. нечувствителен к точному значению коэффициента усиления).

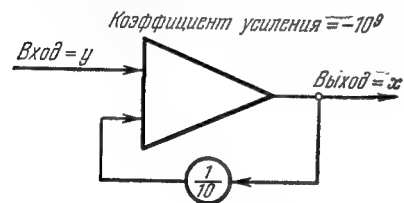


Рис. 13.1-1.

При некоторых обстоятельствах, если существует достаточная задержка при возвращении сигнала с выхода на вход, система осциллирует или, как часто говорят, «шумит». Такие ситуации часто возникают в пу-

бличных аудиториях, которые имеют микрофон (y) и выход усилительной системы (x). Выход из громкоговорителя задерживается на время, которое требуется звуку, чтобы добраться от громкоговорителя до микрофона, и если не позаботиться о регулировке, то результат будет слышен аудитории в виде шума.

Теория систем с обратной связью хорошо развита и не может быть изложена здесь подробно. Следует лишь уяснить, что когда результат некоего процесса используется в качестве входа в него же на более поздней стадии, то есть риск получить нежелательные колебания, период и режим которых зависят от данного процесса.

Второй новой идеей является развитие приема, которым мы пользовались для вывода формул. До сих пор использовались все имеющиеся параметры, чтобы сделать формулу верной для возможно большего числа степеней x . Другими словами, мы стремились сделать порядок остаточного члена возможно более высоким. Как оказывается, чтобы справиться с неустойчивостью и другими эффектами, такими как, например, накопление ошибок округления, необходимо оставить некоторые коэффициенты общей формулы неопределенными до нахождения остаточного члена и лишь тогда выбрать для них значения, которые окончательно и дадут нужную формулу.

Вычисление неопределенного интеграла

$$y(x) = \int_a^x f(x) dx \quad (13.1-1)$$

эквивалентно решению дифференциального уравнения

$$y'(x) = f(x), \quad y(a) = 0. \quad (13.1-2)$$

Значения подинтегральной функции будем обозначать через y' , а ответ — через y . Эти обозначения хорошо согласуются с системой обозначений, принятой при рассмотрении дифференциальных уравнений более общего вида

$$y'(x) = f(x, y), \quad y(a) = b \quad (13.1-3)$$

(см. гл. 15). Естественно, что при этом класс формул, среди которых следует искать нужную, сужается. Сделаем также незначительные изменения в обозначениях. До сих пор $y(x)$ и y_x использовались для обозначения одного и того же. Теперь мы будем применять индексные обозначения; точки, в которых мы вычисляем ответ, будем нумеровать 0, 1, 2, ..., шаг по-прежнему обозначается через h , а начало отсчета — через a . Таким образом, $y(a + nh)$ и y_n есть одна и та же величина. Путаницы при этом не возникает, а обозначения становятся значительно проще.

§ 13.2. Несколько простых формул для неопределенных интегралов

Вероятно, самая простая формула аппроксимации одного шага неопределенного интеграла такая:

$$y_{n+1} = y_n + hy' \left[a + \left(n + \frac{1}{2} \right) h \right] = y_n + hy'_{n+1/2}, \quad (13.2-1)$$

где

$$y' = f(x)$$

или

$$y = \int_a^x y' dx.$$

Очевидно, это — формула средней точки (10.2-1) и она дает ошибку вдвое меньше, чем обычное правило трапеций (10.2-3)

$$y_{n+1} = y_n + \frac{h}{2} (y'_{n+1} + y'_n).$$

Формула средней точки лучше также с точки зрения эффекта округления.

Более точная формула, которая может быть найдена при использовании универсальной матрицы § 10.3, есть

$$y_{n+1} = y_n + \frac{h}{24} (-y'_{n-1} + 13y'_n + 13y'_{n+1} - y'_{n+2}). \quad (13.2-2)$$

Ее остаточный член равен

$$\frac{11}{720} h^5 y^{(5)}(\theta) \approx 0,0153 h^5 y^{(5)}(\theta). \quad (13.2-3)$$

Если для вычисления интеграла используется эта формула, то нужно сосчитать функцию в двух дополнительных точках, лежащих вне области интегрирования (по одной на каждом конце). Если значения подынтегральной функции не могут быть найдены в этих точках из-за особенностей $f(x)$, то предположение, что подынтегральная функция может быть хорошо аппроксимирована многочленом, вероятно, неверно.

Для концов интервала можно использовать другую формулу, а именно:

$$y_1 = y_0 + \frac{h}{24} (9y'_0 + 19y'_1 - 5y'_2 + y'_3)^*. \quad (13.2-4)$$

Для вычисления неопределенных интегралов широко используется формула Симпсона (10.4-1). В наших теперешних обозначениях она принимает вид

$$y_{n+1} = y_{n-1} + \frac{h}{3} (y'_{n-1} + 4y'_n + y'_{n+1}), \quad (13.2-5)$$

с остаточным членом

$$-\frac{h^5 y^{(5)}(\theta)}{90} \approx -0,0111 h^5 y^{(5)}(\theta), \quad (13.2-6)$$

который несколько меньше, чем (13.2-3).

Когда формула Симпсона используется стандартным образом, возникает одна трудность, которая вообще не очень серьезна, но при случае может досадить. Формула Симпсона дает значение интеграла на два шага вперед, а чтобы начать цепочку для нечетных узлов, используется специальная «половинная формула Симпсона» (10.4-2). Результатом этого прыжка на два шага вперед является то, что накоп-

*) На самом деле мы, конечно, вычисляем $\frac{24}{h} y_n$ и только в конце счета, когда печатаем y_n , умножаем на $\frac{h}{24}$. Поэтому при том же количестве вычислений мы избегаем накопления ошибок округления.

ливающиеся ошибки по нечетным и четным точкам до некоторой степени не зависят друг от друга, особенно из-за разницы в весах, с которыми один и тот же узел входит в четные и нечетные последовательности. Все это способствует возникновению колебаний. Хотя эти колебания происходят вследствие сделанных ошибок и поэтому дают некоторую меру точности результатов, иногда они могут оказаться очень неприятными. Колебаний можно избежать, если вычислять лишь ценочку четных точек, а для каждой нечетной точки использовать «половинную формулу Симпсона».

Упражнения

13.2-1. Вывести формулы (13.2-2) и (13.2-4) с их остаточными членами.

13.2-2. Сравнить метод формулы Симпсона с формулой соответствующей точности, основанной на формуле Грегори (§ 10.10).

§ 13.3. Общий метод

Можно исследовать много специальных формул в отдельности, но мы вернемся к общему методу и исследуем целый класс сразу. При приближенном вычислении следующего значения интеграла y_{n+1} можно было бы пользоваться старыми значениями интеграла y_n , y_{n-1} , а также значениями подынтегральной функции y'_{n-1} , y_n , y'_{n+1} , ... Здесь мы сознательно ограничимся формулой вида

$$y_{n+1} = a_0 y_0 + a_1 y_{n-1} + a_2 y_{n-2} + h(b_{-1} y'_{n+1} + b_0 y'_n + b_1 y'_{n-1} + b_2 y'_{n-2}) + E_5 \frac{h^5 y^{(5)}(\theta)}{5!}, \quad (13.3-1)$$

которая использует три старых значения интеграла и одно новое и три старых значения подынтегральной функции.

Немедленно возникает вопрос: «почему бы не попробовать более разумный метод, использующий значения подынтегральной функции в точках, расположенных симметрично относительно той, значение в которой надо вычислить»? Такой метод, вероятно, дал бы более точные формулы, но мы не будем так делать по причинам, которые выяснятся при исследовании общего дифференциального уравнения (13.1-3) в гл. 15.

На семь параметров в формуле (13.3-1) наложим пока лишь пять условий, а именно, что формула должна быть точной для $y=1$, x, \dots, x^n или, что то же самое, точной для $y=1$, $(x-h), \dots, (x-h)^4$, и сохраним два параметра для других целей. Раньше 1, x , x^2, \dots были подынтегральными функциями, теперь же под $y=1$, x , x^2, \dots мы понимаем ответ. Но интеграл от степени x есть степень x на единицу большая, поэтому нужно еще объяснить, зачем мы прибавили условие точности для $y=1$. Здесь мы не дадим объяснения этому, однако укажем, что невыполнение этого требования влечет нежелательные следствия.

Используя указанные пять условий и взяв a_1 и a_2 в качестве параметров, получим

$$\left. \begin{aligned} a_0 &= 1 - a_1 - a_2, & b_0 &= \frac{1}{24}(19 + 13a_1 + 8a_2), \\ a_1 &= a_1, & b_1 &= \frac{1}{24}(-5 + 13a_1 + 32a_2), \\ a_2 &= a_2, & b_2 &= \frac{1}{24}(1 - a_1 + 8a_2), \\ b_{-1} &= \frac{1}{24}(9 - a_1), & E_3 &= \frac{1}{6}(-19 + 11a_1 - 8a_2). \end{aligned} \right\} \quad (13.3-2)$$

Мы вычислили E_3 (и другие коэффициенты) так, как если бы шаг h был равен единице. Разбор одного примера с разбиением h покажет читателю, почему было решено, что такой способ нахождения коэффициентов (который действительно несколько меняет смысл E_3) предпочтительнее на практике. Это не приводит к путанице в понимании E_3 .

Упражнение 13.3-1. Вывести уравнения (13.3-2)

§ 13.4. Ошибка вследствие отбрасывания членов

Общий вид формулы (13.3-1) написан так, как если бы функция влияния $G(s)$ (см. § 11.3) имела постоянный знак (§ 11.4). Необходимо ее исследовать и найти, какие значения a_1 , a_2 делают это предположение верным.

Руководствуясь гл. 11, находим, во-первых, значение $m=5$ (за исключением специальных a_1 , a_2 (см. § 11.2)). Из функции влияния § 11.3 находим $A = -2h$, так что формула (11.3-2) приобретает вид

$$\begin{aligned} y(x) &= y(-2h) + (x+2h)y'(-2h) + \frac{(x+2h)^2}{2!}y''(-2h) + \\ &+ \frac{(x+2h)^3}{3!}y'''(-2h) + \frac{(x+2h)^4}{4!}y^{(4)}(-2h) + \\ &+ \frac{1}{4!} \int_{-2h}^x y^{(5)}(s)(x-s)^4 ds. \end{aligned} \quad (13.4-1)$$

Если подставить это выражение в общую формулу (13.3-1), то окончательно придем к формуле (11.3-8) с

$$R(y) = \int_{-2h}^h y^{(5)}(s) G(s) ds,$$

где $G(s)$ задана следующим равенством (заметим, что для получения этого равенства не надо интегрировать):

$$G(s) = \frac{1}{4!} \{ (h-s)_+^4 - a_0(-s)_+^4 - a_1(-h-s)_+^4 - a_2(-2h-s)_+^4 - \\ - 4h[b_{-1}(h-s)_+^3 + b_0(-s)_+^3 + b_1(-h-s)_+^3 + \\ + b_2(-2h-s)_+^3] \}. \quad (13.4-2)$$

Если $G(s)$ имеет постоянный знак (§ 11.4), то

$$R(x^5) = 5! \int_{-2h}^h G(s) ds = E_5 h^5,$$

где E_5 есть результат подстановки $y(x) = x^5$ для разбиения с $h = 1$. Задача состоит в том, чтобы найти такие значения a_1, a_2 , при которых $G(s)$ имеет постоянный знак. Таким образом, нам надо найти нули $G(s)$ для каждой пары a_1, a_2 . Так как это сделать трудно, прибегнем к обычному приему и обратим задачу: положим $G(s) = 0$ и исследуем полученные в результате значения a_1, a_2 .

Используя (13.3-2), напомним $G(s)$ как линейную функцию от a_1, a_2 :

$$G(s) = G_0(s) + a_1 G_1(s) + a_2 G_2(s),$$

где

$$G_0(s) = \frac{1}{4!} \left[(h-s)_+^4 - (-s)_+^4 - \frac{3}{2} h(h-s)_+^3 - \frac{19}{6} h(-s)_+^3 + \right. \\ \left. + \frac{5}{6} h(-h-s)_+^3 - \frac{h}{6} (-2h-s)_+^3 \right],$$

$$G_1(s) = \frac{1}{4!} \left[(-s)_+^4 - (-h-s)_+^4 + \frac{h}{6} (h-s)_+^3 - \frac{13}{6} h(-s)_+^3 - \right. \\ \left. - \frac{13}{6} h(-h-s)_+^3 + \frac{h}{6} (-2h-s)_+^3 \right],$$

$$G_2(s) = \frac{1}{4!} \left[(-s)_+^4 - (-2h-s)_+^4 - \frac{4}{3} h(-s)_+^3 - \frac{16}{3} (-h-s)_+^3 - \right. \\ \left. - \frac{4}{3} h(-2h-s)_+^3 \right].$$

Мы должны исследовать $G(s)$ в следующих трех интервалах:

$$h \geq s \geq 0, \quad 0 \geq s \geq -h \quad \text{и} \quad -h \geq s \geq -2h.$$

В первом интервале равенство

$$(h-s)^3 \left[(h-s) - \frac{3}{2} h + a_1 \frac{h}{6} \right] = 0$$

дает

$$a_1 = 3 + \frac{6s}{h},$$

что описывает семейство вертикальных прямых в плоскости (a_1, a_2) . Прямые двигаются от $a_1=9$ до $a_1=3$, когда s проходит от h до 0 (рис. 13.4-1).

Во втором интервале имеем

$$\left[(h-s)^4 - s^4 - \frac{3}{2} h (h-s)^3 + \frac{19}{6} h s^3 \right] + \\ + a_1 \left[s^4 + \frac{h}{6} (h-s)^3 + \frac{13}{6} h s^3 \right] + a_2 \left[s^4 + \frac{4h}{3} s^3 \right] = 0.$$

Это уравнение описывает прямые, которые продолжают двигаться влево, но постепенно наклоняются до тех пор, пока при $s = -h$

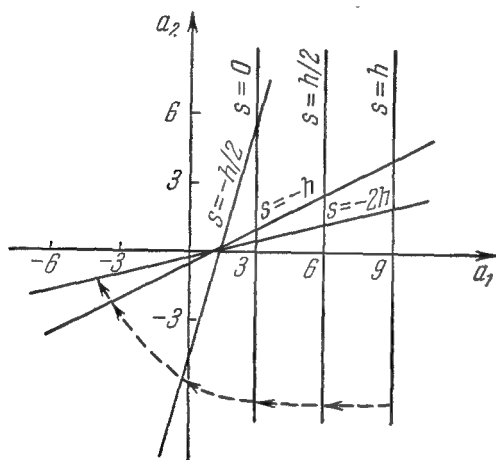


Рис. 13.4-1. Область постоянства знака $G(s)$ (прямые — линии уровня $G(s)=0$ для указанных значений s).

сохраняет постоянный знак. Таким образом, в этой области остаточный член имеет вид

$$\frac{E_s h^5}{5!} y^{(5)}(\theta).$$

Вне этого района мы не имеем такого остаточного члена. Дальше этот вопрос исследоваться не будет.

В этой области ошибка выражается через E_s согласно (13.3-2). Прямые «постоянной ошибки» изображены на рис. 13.4-2 светлыми

$$a_2 = \frac{a_1 - 1}{2}.$$

Эта прямая проходит через точку $(1, 0)$ с наклоном 1:2.

В третьем интервале получаем прямую, вращающуюся около точки $(1, 0)$ до положения

$$a_2 = \frac{a_1 - 1}{4}.$$

Выше и слева от этой прямой (см. рис. 13.4-1) функция влияния $G(s)$ не имеет нулей*) и, следовательно,

*) Существует также область справа внизу, в которой $G(s)$ не меняет знака, но как величина остаточного члена, так и эффекты округления удерживают нас от использования таких формул; в дальнейшем мы будем игнорировать эту область.

линиями. Эти линии показывают, что при прочих равных условиях мы должны попытаться держаться ниже и правее, где остаточный член

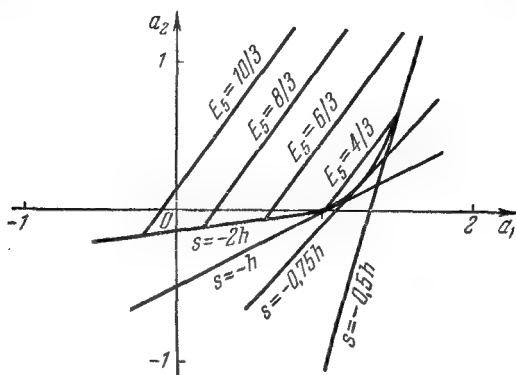


Рис. 13.4-2. Линии равной ошибки (E — значение ошибки вдоль каждой линии).

маленький. Наклон линий 11:8 показывает изменение отношения двух параметров; если увеличивать a_2 на 11 единиц, а a_1 на 8, то ошибка остается той же самой.

§ 13.5. Устойчивость

Исследуем сначала полученные формулы на тривиальном случае

$$y' = 0. \quad (13.5-1)$$

Если уже здесь возникнут затруднения, то в более общем случае

$$y' = f(x) \quad (13.5-2)$$

их следует ожидать и подалее.

Если в качестве точки плоскости (a_1, a_2) выбрать $a_1 = -1$, $a_2 = 0$, то формула примет вид

$$y_{n+1} = 2y_n - y_{n-1}.$$

Рассмотрим, что случится с решением $y=0$, если допустить в y_0 небольшую ошибку ϵ (представим себе, что она произошла от округления). Легко вычислить таблицу 13.5-1. Таким образом, ошибка все время возрастает.

В качестве другого примера выберем $a_1 = 1$, $a_2 = 1$. Уравнение есть

$$y_{n+1} = -y_n + y_{n-1} + y_{n-2}$$

и получается таблица 13.5-2. Теперь ошибка раскачивается и также растет.

Т а б л и ц а 13.5-1

.....	
$y_{-2} = 0$	$y_1 = 2\varepsilon$
$y_{-1} = 0$	$y_2 = 3\varepsilon$
$y_0 = \varepsilon$	$y_3 = 4\varepsilon$

Т а б л и ц а 13.5-2

.....	
$y_{-2} = 0$	$y_2 = 2\varepsilon$
$y_{-1} = 0$	$y_3 = -2\varepsilon$
$y_0 = \varepsilon$	$y_4 = 3\varepsilon$
$y_1 = -\varepsilon$	$y_5 = -3\varepsilon$

Эти два примера показывают, что это явление надо исследовать подробнее; ясно, что некоторые точки (a_1, a_2) очень плохо выбраны.

Общая формула (13.3-1), которая исследуется, является линейным разностным уравнением с постоянными коэффициентами, и можно применить методы гл. 5. Так как значения y' есть значения подинтегральной функции, которые вычисляются независимо от значений y , то рассмотрим разностное уравнение в виде

$$y_{n+1} - a_0 y_n - a_1 y_{n-1} - a_2 y_{n-2} = F_n. \quad (13.5-3)$$

Прежде всего, решим однородное уравнение

$$y_{n+1} - (1 - a_1 - a_2) y_n - a_1 y_{n-1} - a_2 y_{n-2} = 0. \quad (13.5-4)$$

Для этого положим $y_n = \rho^n$ и получим характеристическое уравнение

$$\rho^3 - (1 - a_1 - a_2) \rho^2 - a_1 \rho - a_2 = 0$$

или

$$(\rho - 1)[\rho^2 + (a_1 + a_2)\rho + a_2] = 0. \quad (13.5-5)$$

Пусть ρ_1 и ρ_2 — нули квадратичного множителя. Решение (13.5-4) при условии, что никакие два корня не совпадают, есть

$$y_n = C_1(\rho_1)^n + C_2(\rho_2)^n + C_3. \quad (13.5-6)$$

Когда два корня равны, как это случилось в двух верхних примерах, формулу (13.5-6) надо изменить. Если каждый из двух корней

в (13.5-6) по модулю больше единицы, то решение будет расти подобно геометрической прогрессии по мере того, как n увеличивается (при условии, что соответствующие коэффициенты не нули).

Поэтому исследуем в плоскости (a_1, a_2) область, для которой $|\rho_i| < 1$. Самый легкий путь — определить границы этой области.

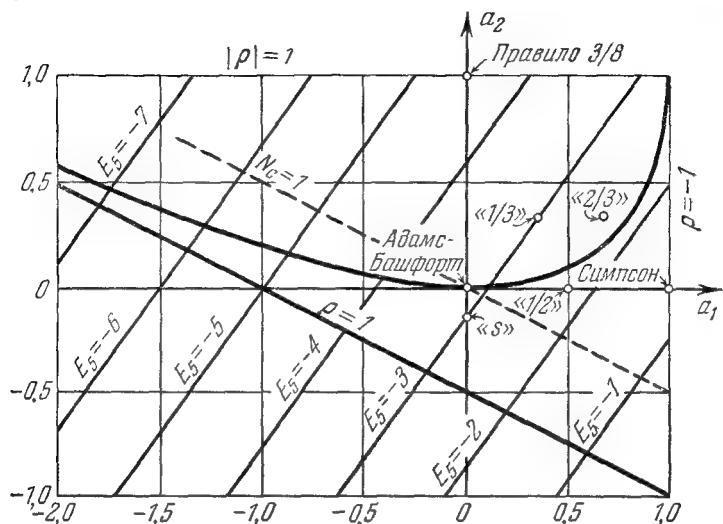


Рис. 13.5-1. Область устойчивости на плоскости.

Сначала исследуем границу, вдоль которой корень становится равным 1. Для этого положим $\rho = 1$ в квадратичном множителе. Это определяет прямую

$$1 + a_1 + 2a_2 = 0, \quad a_2 = -\frac{1+a_1}{2}. \quad (13.5-7)$$

Далее, исследуем, где $\rho = -1$; положим в квадратичном множителе $\rho = -1$. Это определит прямую

$$a_1 = 1. \quad (13.5-8)$$

Теперь исследуем, где корни становятся комплексными. Граничная кривая определяется уравнением

$$(a_1 + a_2)^2 - 4a_2 = 0$$

или

$$a_1 = \pm 2\sqrt{a_2} - a_2 \quad (13.5-9)$$

(эта кривая показана на рис. 13.5-1). Внутри параболы корни комплексные, взаимно сопряженные и оба имеют модуль

$$|\rho| = \sqrt{a_2}.$$

Таким образом, область, в которой $|p_i| < 1$, ограничена прямой

$$a_3 = 1, \quad (13.5-10)$$

которая как раз пересекает параболу (13.5-9) там, где она касается прямых (13.5-7) и (13.5-8). Эти три прямые образуют треугольник (рис. 13.5-1), внутри которого и должны лежать характеристические корни, чтобы метод интегрирования имел ошибку, которая не растет как геометрическая прогрессия. Вне этой области некоторые $|p_i| > 1$. Легко видеть, что два иллюстрирующих примера здесь соответствуют точкам, которые лежат на краю области и дают двойные корни в характеристическом уравнении, а тройной корень в свою очередь приводит к линейному росту. Когда $|p_i| < 1$, мы говорим, что метод устойчивый; когда $|p_i| = 1$, мы говорим, что он условно устойчивый (лишь для простых корней).

Решение однородного уравнения (13.5-4) должно быть добавлено к решению полного уравнения (13.5-3). Когда решается (13.5-3), любая ошибка округления будет начальным условием для решения однородного уравнения и породит решение, которое стремится к бесконечности (если только не окажется, что p_i по модулю не больше единицы) независимо от поведения F_n .

Нежелательные решения возникают оттого, что дифференциальное уравнение первого порядка заменяется разностным уравнением третьего порядка. Корень $p=1$ соответствует желаемому решению, а другие два p_i — нежелательным решениям. При требовании $|p_i| \leq 1$ они не оказывают большого влияния на окончательный ответ.

Упражнения

13.5-1. Рассмотрите пример $a_1 = -1$, $a_2 = 0$ и покажите, используя уравнение (13.5-6), что вычисленная таблица совпадает с теоретической.

13.5-2. То же самое, что и в упражнении 13.5-1, для $a_1 = 1$, $a_2 = 1$.

13.5-3. Рассмотреть случай $a_1 = 2$, $a_2 = 0$. Показать, что, в соответствии с теорией, ошибка растет в геометрической прогрессии.

13.5-4. Рассмотреть поведение ошибки при пересечении каждой из трех граничных линий области устойчивости.

§ 13.6. Шум округления

Значения y_n , которые мы вычисляем, будут, вообще говоря, иметь некоторый уровень шума округления σ . На первый взгляд кажется, что коэффициент усиления шума округления равен

$$N_a = (a_0^2 + a_1^2 + a_2^2)^{1/2}.$$

Однако, так как значение y_{n+1} вычисляется из значений y_n , y_{n-1} и y_{n-2} , шум в y_{n+1} связан (коррелирован) с шумом в y_n , y_{n-1} , y_{n-2} ; поэтому следует отнестись к вопросу более внимательно.

Рассмотрим решение уравнения

$$y' = 0,$$

которое есть тождественный нуль; предположим, что возникла маленькая ошибка округления. Решение (13.5-6) есть

$$y_n = C_1 (\rho_1)^n + C_2 (\rho_2)^n + C_3. \quad (13.6-1)$$

Если метод устойчив и $|\rho_i| < 1$, то

$$y_n \rightarrow C_3 \quad \text{когда } n \rightarrow \infty.$$

Вычислив C_3 из начальных данных

$$y_{-2} = 0, \quad y_{-1} = 0, \quad y_0 = \varepsilon,$$

имеем

$$\frac{C_1}{\rho_1^2} + \frac{C_2}{\rho_2^2} + C_3 = 0, \quad \frac{C_1}{\rho_1} + \frac{C_2}{\rho_2} + C_3 = 0, \quad C_1 + C_2 + C_3 = \varepsilon.$$

Решая систему относительно C_3 , получаем

$$C_3 = \frac{\varepsilon}{1 - (\rho_1 + \rho_2) + \rho_1 \rho_2}.$$

Используя квадратичный множитель в (13.5-5), находим

$$\rho_1 + \rho_2 = -(a_1 + a_2), \quad \rho_1 \rho_2 = a_2.$$

Следовательно,

$$C_3 = \frac{\varepsilon}{1 + a_1 + 2a_2} = N_c \varepsilon, \quad (13.6-2)$$

где

$$N_c = \frac{1}{1 + a_1 + 2a_2}.$$

Таким образом, чтобы ошибка округления ε не росла, надо чтобы

$$|N_c| = |1 + a_1 + 2a_2|^{-1} \leq 1$$

и чем меньше, тем лучше. Так как обычно приходится иметь дело с положительным выражением $1 + a_1 + 2a_2$, то условие, чтобы изолированная ошибка ε не росла, состоит в том, чтобы точка (a_1, a_2) лежала выше прямой

$$a_2 = -\frac{a_1}{2}. \quad (13.6-3)$$

Эта область ограничена на рис. 13.5-1 пунктирной линией.

Упражнение 13.6-1. Теория округления, приводящая к уравнению (13.6-3), была развита для единственной изолированной ошибки. Однако каждый шаг будет порождать некоторую ошибку. Как изменить аргументацию применительно к этому случаю?

§ 13.7 Итоги

Таблица 13.7-1 перечисляет ряд хорошо известных методов, которые попадают в общий класс (уравнение (13.3-1)), исследуемый нами.

Таблица 13.7-1

Некоторые методы интегрирования

	Адамс — Башфорт	Симпсон	Правило трех вось- мых	1/3	1/2	2/3	«S»
a_0	1	0	0	1/3	1/2	0	9/8
a_1	0	1	0	1/3	1/2	2/3	0
a_2	0	0	1	1/3	0	1/3	-1/8
b_1	9/24	1/3	3/8	13/36	17/48	25/72	3/8
b_0	19/24	4/3	9/8	39/36	51/48	91/72	6/8
b_1	-5/24	1/3	9/8	15/36	3/48	43/72	-3/8
b_2	1/24	0	3/8	5/36	1/48	9/72	0
E_5	-19/6	-4/3	-9/2	-3	-9/4	43/18	-3
Усиление шума N_c	1	1, 0, 1, 0	1, 0, 0, 1, 0, 0	1/2	2/3	3/7	1, 13
N_a	1	1	1	1/√3	1/√2	√5/3	√82/8

Точка $a_1=1$, $a_2=0$ дает метод Симпсона и расположена на границе области устойчивости.

Точка $a_1=0$, $a_2=1$ дает правило трех восьмых (12.2-3) и также лежит на границе области устойчивости.

Точка $a_1=0$, $a_2=0$ дает формулу «Адамса — Башфорта», названную так потому, что она использована в методе Адамса — Башфорта интегрирования дифференциальных уравнений (см. гл. 15).

Все другие методы общего вида (13.3-1) являются линейными комбинациями этих трех. Рис. 13.5-1 дает соответствующую информацию. Каждая точка в плоскости (a_1, a_2) соответствует методу интегрирования.

Во-первых, мы стремимся выбрать точку внутри треугольника устойчивости, хотя можно осмелиться в некоторых случаях выйти на границу, как в методе Симпсона и правиле трех восьмых. Во-вторых, следует держаться направления вниз направо, чтобы иметь возможно меньшее E_5 (рис. 13.4-2). В-третьих, мы хотим держаться выше пунктирной линии (рис. 13.5-1), чтобы сохранить как можно меньший шум.

Таким образом, за исключением беспокоящей тенденции к колебанию, формула Симпсона — одна из лучших в рассмотренной общей форме (которая не включает (13.2-2)). Потребителю следует отобрать себе собственный метод, подходящий для конкретного соотношения

факторов в каждой проблеме; нельзя предложить единственную, самую хорошую формулу при всех обстоятельствах. Методы из таблицы 13.7-1, названные $1/3$, $1/2$ и $2/3$, являются смесью основных методов и интересны в некоторых приложениях.

Упражнение 13.7-1. Развить теорию § 13.3 до § 13.7 (используя только один параметр a_1) для формул общего вида

$$y_{n+1} = a_0 y_n + a_1 y_{n-1} + h(b_{-1} y'_{n-1} + b_0 y'_n + b_1 y'_{n+1}) + E_4 \frac{h^4 y^{(IV)}(\theta)}{4!}$$

и сравнить с известными результатами.

О т в е т:

	$a_1 = 0$	Симпсон	$a_1 = 1/2$		$a_1 = 0$	Симпсон	$a_1 = 1/2$
$a_0 = 1 - a_1$	1	0	$1/2$	$b_1 = \frac{-1 + 5a_1}{12}$	$-1/12$	$1/3$	$1/8$
$a_1 = a_1$	0	1	$1/2$	$E_4 = -1 + a_1$	-1	0	$-1/2$
$b_{-1} = \frac{5 - a_1}{12}$	$5/12$	$1/3$	$3/8$	Усиление шума			
$b_0 = \frac{8 + 8a_1}{12}$	$8/12$	$4/3$	$8/8$	$N_c = \frac{1}{1 + a_1}$	1	1, 0, 1, 0, ..., 1	$2/3$
				N_a	1	1	$1/\sqrt{2}$

$\rho_1 = -a_1$, следовательно, область устойчивости $1 < a_1 < 1$.

§ 13.8. Некоторые общие замечания

Внимательный читатель, вероятно, заметил, что был исследован лишь один особый класс формул; но возможны и другие формулы. Например, можно рассматривать формулы, в которых использовались бы как значение подынтегральной функции (y'), так и значение производной подынтегральной функции (y''). В этом классе естественно ожидать более хороших формул в смысле эффективности вычислений, так как часто производная легко вычисляется из различных кусков вычисления подынтегральной функции. Это потребовало бы большей кодировки (больше ошибок) и для случайной задачи, возможно, было бы плохой стратегией. Однако для часто повторяющейся задачи такие формулы могли бы значительно сократить машинное время и должны быть исследованы.

Мощные формулы типа Гаусса не приспособлены для неопределенного интеграла из-за трудности подгонки специфического разбиения к последовательным шагам.

Применение формул средней точки (13.2-1) (значения y' в средних точках, значения y в узловых), насколько известно автору, недостаточно исследовано. Оно также должно дать интересные результаты.

Выбор E_3 в качестве остаточного члена нашел широкое распространение в практике и учитывает смесь различных факторов. Специальные случаи, конечно, требуют другого выбора.

Главной целью этой главы было ввести идею устойчивости и метод выделения параметров для ее достижения. Поэтому мы не вдавались в сложный вопрос о том, как начать интегрирование, пока нет предыдущих значений, стоящих в правой части общих формул. Иногда для этой цели могут быть использованы формулы § 13.2.

Упражнения

13.8-1. Исследовать класс формул

$$y_{n+1} = y_n + h(b_{-1}y'_n + 1 + b_0y'_n) + h^2(c_{-1}y''_n + 1 + c_0y''_n) + E_4 \frac{h^4 y^{(IV)}(\theta)}{4!},$$

используя b_{-1} в качестве параметра. Обратить внимание на начало.

Ответ:

	Общее решение		Метод Мил- ля *)			
b_{-1}	b_{-1}	1	1/2	0	1/3	2/3
b_0	$1 - b_{-1}$	0	1/2	1	2/3	1/3
c_{-1}	$\frac{1 - 3b_{-1}}{6}$	-2/6	-1/12	1/6	0	-1/6
c_0	$\frac{2 - 3b_{-1}}{6}$	-1/6	1/12	1/6	1/6	0
E_4	$-1 + 2b_{-1}$	1	0	-1	-1/3	1/3
E_5	1/6			

*) См. [26].

$G_4(s)$ меняет знак для $1/3 < b_{-1} < 2/3$, но в случае $b_{-1} = 1/2$ ошибка

$$\frac{E_5}{5!} h^5 y^{(5)}(\theta) = \frac{h^5}{720} y^{(5)}(\theta),$$

так как

$$G(s) = \frac{(h-s)^2 s^2}{4!} \neq 0 \quad (0 < s < h),$$

и этот случай особенно предпочтителен.

13.8-2. Развить метод для начала интегрирования в общем виде (13.3-1) с должной заботой о точности.

§ 13.9. Экспериментальная проверка устойчивости

Читателю не следует пугаться разработанной выше сложной теории. Если надо исследовать обширный класс формул, необходим примерно такой метод, но если рассматривается отдельная формула, то часто удовлетворяются ее экспериментальной проверкой.

Предположим, что рассматривается формула

$$y_{n+1} = -y_n + \frac{3}{2}y_{n-1} + \frac{y_{n-2}}{2} + \frac{h}{48}(15y'_{n+1} + 85y'_n + 61y'_{n-1} + 7y'_{n-2}).$$

Легко увидеть, подставив $y=1$, x , x^2 , x^3 , x^4 , удовлетворяют ли они точно этому уравнению.

Что касается устойчивости, то просто берут

$$y_{-2} = 0, \quad y_{-1} = 0, \quad y_0 = \varepsilon$$

и вычисляют

$$y_1 = -\varepsilon, \quad y_2 = \frac{5}{2}\varepsilon, \quad y_3 = -\frac{7}{2}\varepsilon, \quad \dots$$

В данном случае ясно, что метод неустойчив. Чтобы иметь уверенность, что эксперименты охватывают все возможные начальные ситуации, нужно проявить некоторое внимание (испытать каждое $c_i \neq 0$ хотя бы раз). Обычно конкретное испытание начинается с линейной комбинации всех таких членов.

Простая экспериментальная проверка того, что метод нестабилен, убедительна и легко применима. Обратное заключение относительно устойчивости требует большего внимания.

Этот метод иллюстрируется в следующем параграфе.

§ 13.10. Пример интеграла свертки, иллюстрирующий идею устойчивости

Колебания с геометрическим ростом амплитуды являются типичными для неполадок с устойчивостью. Следующий пример показывает, как различные идеи могут применяться в конкретных случаях.

Задача состояла в том, чтобы найти $\Phi(t)$, если $\psi(t)$ была задана в точках t_i ($t_i = 10^{(-14.4 + i0.2)}$, $i=0, 1, 2$), уравнение, связывающее $\Phi(t)$ и $\psi(t)$, было

$$\int_0^t \Phi(t-\tau) \psi(\tau) d\tau = t \quad (13.10-1)$$

или эквивалентное

$$\int_0^t \Phi(\tau) \psi(t-\tau) d\tau = t \quad (13.10-2)$$

причем $\psi(t)$ была монотонно возрастающей.

Был выбран и испытан метод вычислений, выглядевший разумным. Результаты расчетов дали колебания с растущей амплитудой. Период колебаний был фиксированным, а рост амплитуды — более или менее геометрическим; отсюда было сделано предположение, что это — следствие неустойчивости. Исследование процесса вычисления проводилось для $\psi(t) = 1$, но с единственным возмущением; такие данные приводили к колебаниям.

После некоторого изучения была принята следующая новая схема вычисления. Мы положили

$$f(t) = \int_0^t \psi(\tau) d\tau.$$

Таким образом, $f(0) = 0$ и $f'(\tau) = \psi(\tau)$. Далее ввели обозначения t_i ($i = 0, 1, \dots, n$) с $t_0 = 0$, $t_1 = 10^{-14.4}$, ..., $t_i = 10^{(-14.6 + 0.2i)}$ и допустили, что $\psi(t)$ между t_0 и t_1 является константой. Чтобы вычислить $f(t)$, применили для интегрирования правило трапеций

$$f(t_{n+1}) = f(t_n) + \frac{1}{2} [\psi(t_{n+1}) + \psi(t_n)](t_{n+1} - t_n),$$

так как данные не оправдывали сколько-нибудь более сложного метода.

Теперь из (13.10-2) получается

$$t_{n+1} = \int_0^{t_{n+1}} \Phi(\tau) \psi(t_{n+1} - \tau) d\tau = \sum_{i=0}^n \int_{t_i}^{t_{i+1}} \Phi(\tau) \psi(t_{n+1} - \tau) d\tau.$$

В каждом из интегралов в сумме можно сделать аппроксимацию

$$\begin{aligned} \int_{t_i}^{t_{i+1}} \Phi(\tau) \psi(t_{n+1} - \tau) d\tau &= \Phi(t_{i+1/2}) \int_{t_i}^{t_{i+1}} f'(t_{n+1} - \tau) d\tau = \\ &= -\Phi(t_{i+1/2}) [f(t_{n+1} - t_{i+1}) - f(t_{n+1} - t_i)], \end{aligned}$$

где $t_{i+1/2}$ есть среднее значение $\frac{1}{2}(t_{i+1} + t_i)$. Таким образом,

$$\begin{aligned} t_{n+1} &= - \sum_{i=0}^{n-1} \Phi(t_{i+1/2}) [f(t_{n+1} - t_{i+1}) - f(t_{n+1} - t_i)] + \\ &\quad + \Phi(t_{n+1/2}) f(t_{n+1} - t_n). \end{aligned}$$

Решая это уравнение относительно $\Phi(t_{n+1/2})$, получили

$$\Phi(t_{n+1/2}) = \frac{t_{n+1} - \sum_{i=0}^{n-1} \Phi(t_{i+1/2}) [f(t_{n+1} - t_i) - f(t_{n+1} - t_{i+1})]}{f(t_{n+1} - t_n)}. \quad (13.10-3)$$

Это дает метод для последовательного вычисления каждого значения $\Phi(t_{n+1/2})$, где первое значение

$$\Phi_{1/2} = \frac{t_1}{f(t_1)}. \quad (13.10-4)$$

Причина того, что уравнения (13.10-3) и (13.10-4) оказались удачными, а некоторые другие способы аппроксимации — нет, кроется в том, что ошибка в вычислении $\Phi(t)$ (вследствие ошибочных данных или даже просто численного округления произведения) не растет по величине на следующих этапах вычисления. Чтобы наглядно показать это, вычислим коэффициент при $\Phi(t_{n-1/2})$ в выражении для $\Phi(t_{n+1/2})$. Этот коэффициент равен

$$-\frac{f(t_{n+1}-t_{n-1})-f(t_{n+1}-t_n)}{f(t_{n+1}-t_n)-f(t_{n+1}-t_{n+1})},$$

так как $f(0)=0$. Применим теорему о среднем значении к числителю и к знаменателю отдельно:

$$-\frac{\psi(t_{n+1}-\theta_1)(t_n-t_{n-1})}{\psi(t_{n+1}-\theta_2)(t_{n+1}-t_n)} = -\left[\frac{\psi(t_{n+1}-\theta_1)}{\psi(t_{n+1}-\theta_2)}\right]\left[\frac{t_n-t_{n-1}}{t_{n+1}-t_n}\right], \quad (13.10-5)$$

где $t_n > \theta_1 > t_{n-1}$, $t_{n+1} > \theta_2 > t_n$. Каждая из скобок по модулю меньше единицы, так что ошибка в $\Phi(t_{n-1/2})$ умножается на число, меньшее единицы. Таким образом, ошибка колеблется и быстро исчезает.

Так как функции Φ и ψ входили в уравнение симметрично (ср. (13.10-1) и (13.10-2)), то в качестве проверки исходная функция ψ была вычислена тем же методом из полученной функции $\Phi(t)$. Расхождение получилось в пределах экспериментальной ошибки.

Из этого примера видно, что неустойчивость можно исследовать и без характеристического уравнения (ср. (13.5-5)). Однако общая идея обратной связи все же используется. Предыдущее исследование охватило лишь обратную связь от одного члена; на практике обычно бывает необходимо сделать некоторое исследование, чтобы убедиться, что дополнительная обратная связь от других членов не вызывает неустойчивости.

ГЛАВА 14

ВВЕДЕНИЕ В ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ

§ 14.1. Природа и смысл дифференциальных уравнений

Наши современные взгляды на фундаментальные законы природы часто излагаются в виде дифференциальных уравнений. Например, закон Ньютона

$$f = ma = m \frac{d^2x}{dt^2}$$

есть дифференциальное уравнение второго порядка. В качестве другого примера рассмотрим колонию бактерий, растущую в питательной среде. Когда колония изучается в целом, естественно предположить, что в момент t скорость роста пропорциональна имеющемуся количеству $y(t)$, так что

$$y'(t) = ky(t).$$

Таким образом, дифференциальные уравнения часто дают самое простое математическое описание явлений природы.

Если возможно, дифференциальное уравнение решается в точном виде. Учебники дифференциальных уравнений часто оставляют впечатление, что в точном виде можно решить большинство дифференциальных уравнений, но опыт работы не подтверждает этого. Для получения приближенных решений пригодны различные аналитические методы, и перед тем, как обратиться к численному решению, надо исследовать также и их.

Пусть дано дифференциальное уравнение

$$y' = x^2 - y^2. \quad (14.1-1)$$

Что подразумевается под его решением? Оставляя в стороне формальное математическое определение, интуитивно мы имеем в виду кривую $y = y(x)$, в каждой точке (x, y) которой тангенс угла наклона кривой $y'(x)$ дается уравнением (14.1-1). Это — локальное свойство. Решение уравнения означает распространение этого локального свойства на большую область. Конечно, решением является не одна кривая, а скорее всего через любую точку (x_0, y_0) проходит решение уравнения (14.1-1).

§ 14.2. Поле направлений

Идеи предыдущего параграфа могут быть изображены графически. В плоскости (x, y) мы выбираем различные точки и строим тангенс угла наклона, равный $x^2 - y^2$. В каждой такой точке чертим короткую линию с вычисленным тангенсом угла наклона. Эти линии показывают локальное направление решения, и при некотором воображении можно легко нарисовать различные решения при условии, что точки, через которые проводится линия, расположены достаточно часто. Однако часто бывает проще воспользоваться другим методом. Мы называем линии, вдоль которых все короткие штрихи имеют один и тот же наклон, *изоклинами* (линиями одинакового наклона). В рассматриваемом примере находим, что изоклины с тангенсом угла наклона k — гиперболы: $x^2 - y^2 = k$. Поле направлений, изображенное при помощи изоклин, приведено на рис. 14.2-1.

Следует отметить несколько обстоятельств. Если график решения имеет максимум, то он должен лежать на нулевой изоклине. Аналогично точки перегиба можно обнаружить, положив $y'' = 0$:

$$\begin{aligned} y'' = 0 &= 2x - 2yy' = \\ &= 2[x - y(x^2 - y^2)], \\ x &= \frac{1 \pm \sqrt{1 + 4y^4}}{2y}. \end{aligned}$$

Этой кривой на рис. 14.2-1 нет.

Использованием поля направлений, несмотря на его простоту, не следует пренебрегать. Часто оно выявляет природу задачи, так что правильный подход может дать достаточно точное численное решение. В практике автора несколько раз изображение поля направлений сильно помогло решению поставленной задачи, а в одном случае аккуратный чертеж, сделанный на чертежной доске, дал достаточную точность, чтобы полностью ответить на вопрос задачи.

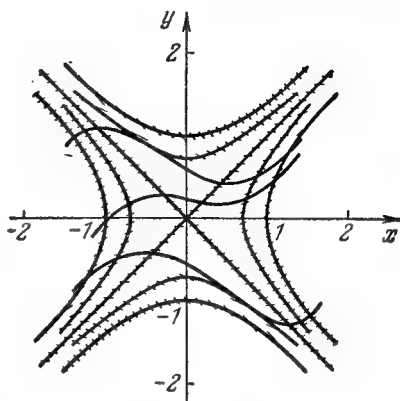


Рис. 14.2-1. Поле направлений для уравнения $y' = x^2 - y^2$.

Упражнения

14.2-1. Начертить поле направлений и решения уравнения $y' = -\sin y$.

14.2-2. Начертить поле направлений и решения уравнения $y' = x - y^2$.

§ 14.3. Численное решение

Если мы хотим получить лишь одно решение, проходящее через данную точку, то было бы бесполезной тратой времени проводить все поле направлений. Скорее следует начать с данной точки, вычислить локальное направление, а затем вычислить тангенсы угла наклона в нескольких близких точках. Так следует продолжать, проводя наклоны лишь там, где по нашим предположениям должно быть решение.

Этот путь может быть сведен к арифметике. Берем данную начальную точку, вычисляем тангенс угла наклона и движемся в этом направлении на небольшое расстояние (смысл слова «небольшое» зависит от конкретной ситуации, но обычно довольно очевиден) и выбираем следующую точку. Используя ее как начальную точку, повторяем процесс сколько нужно.

Рассматривая рис. 14.3-1, легко заметить недостаток этого метода. Если верное решение загибается вверх, то кривая, которую мы вычисляем, всегда отходит вниз от этого решения, потому что для

получения каждой следующей точки используется тангенс угла наклона касательной в предыдущей.

Следовало бы быть умнее: посмотреть предварительно вперед, заметить тангенс угла наклона там, а затем продолжать движение в направлении среднего арифметического начального и конечного тангенсов. Это предложение легко систематизировать. Чтобы узнать, где мы собираемся взять следующую точку, используем формулу

$$y_{n+1} = y_n + 2hy'_n \quad (14.3-1)$$

(Мы берем тангенс угла наклона y'_n в середине удвоенного интервала, чтобы избежать очевидной систематической ошибки.) Ошибка будет равна

$$\frac{h^3}{3} y'''(\theta). \quad (14.3-2)$$

Рис. 14.3-1. Грубое численное решение.

Напишем теперь корректирующую формулу

$$y_{n+1} = y_n + h \frac{y'_{n+1} + y'_n}{2}, \quad (14.3-3)$$

ошибка которой

$$-\frac{h^3}{12} y'''(\theta). \quad (14.3-4)$$

Первый вопрос, который возникает, когда мы пользуемся этой методикой, — с чего начать? Вычисление по первой формуле требует, кроме текущего тангенса угла наклона, знания одной старой точки.

Один из способов получения первой точки состоит в использовании разложения решения в ряд Тейлора относительно начальной точки, коэффициенты которого находятся из уравнения и его производных:

$$y_1 = y_0 + hy'_0 + \frac{h^2}{2!} y''_0 + \frac{h^3}{3!} y'''_0 + \dots \quad (14.3-5)$$

Так как ошибка на каждом шагу зависит от третьей производной, продолжать ряд Тейлора дальше членов с h^3 или h^4 не следует.

Эта схема прогноза и последующего исправления имеет несколько приятных свойств. Разность между предсказанным и исправленным значениями равна

$$\frac{h^3}{12} [4y'''(\theta_1) + y'''(\theta_2)] \approx \frac{5h^3}{12} y'''(\theta). \quad (14.3-6)$$

Если предположить, что y''' не меняет знака в интервале, то эти два значения находятся по разные стороны от верного значения (для

этого шага). Таким образом, метод дает сразу и способ оценки точности. Если ошибка слишком велика, интервал h можно уменьшить до $h' = \frac{h}{2}$, что уменьшает ошибку примерно в восемь раз.

Читатель должен ясно понимать, что мы обращались с нашими символами несколько вольно. Мы оценили новое значение формулой (14.3-1), ошибка которой (14.3-2). Это значение используется для того, чтобы вычислить y'_{n+1} , и, следовательно, хотя мы написали y'_{n+1} как если бы оно было верным, это не так. Ошибка выражается вторым членом формулы

$$f\left(x, y - \frac{h^3}{3} y'''(\theta)\right) \approx f(x, y) + \frac{\partial f(x, y)}{\partial y} \left[-\frac{h^3}{3} y'''(\theta)\right].$$

Обычно сначала строят прогноз, а потом исправляют, и если два полученных значения недостаточно близки, то повторяют исправление несколько раз, используя самое последнее значение, пока изменение не станет достаточно малым. В принципе, лучше сократить интервал, чем производить повторные вычисления по корректирующей формуле.

§ 14.4. Пример

Рассмотрим уравнение

$$y' = y^2 + 1, \quad y(0) = 0 \quad (0 \leq x \leq 1) \quad (14.4-1)$$

(выбранное потому, что решение $y = \operatorname{tg} x$ известно и можно использовать его для проверки).

Прежде всего, найдем несколько первых членов ряда Тейлора (14.3-5), продифференцировав уравнение:

$$\begin{aligned} y' &= y^2 + 1, & y'(0) &= 1, \\ y'' &= 2yy', & y''(0) &= 0, \\ y''' &= 2yy'' + 2(y')^2, & y'''(0) &= 2, \\ y^{IV} &= 2y'y'' + 2yy''' + 4y'y'', & y^{IV}(0) &= 0. \end{aligned}$$

Следовательно,

$$y(x) = 0 + x \cdot 1 + \frac{x^2}{2} \cdot 0 + \frac{x^3}{3!} \cdot 2 + \frac{0 \cdot x^4}{4!} + \dots = x + \frac{x^3}{3} + \dots$$

Используем грубое разбиение $h = \Delta x = 0.2$. В качестве первой точки имеем

$$y_1 = y(0.2) = 0.2 + \frac{0.008}{3} = 0.203$$

и (с помощью логарифмической линейки) находим

$$y'_1 = y(0.2) = (0.203)^2 + 1 = 1.0412.$$

Теперь мы готовы начать построение решения

$$p_2 = y_0 + 2hy'_1 = 0 + 0,4 \cdot (1,0412) = 0,4165$$

и вычисляем, используя уравнение (14.4-1),

$$p'_2 = 1,173.$$

Дальше вычисляем поправочное значение из (14.3-3)

$$c_2 = 0,203 + 0,1 (1,0412 + 1,173) = 0,424.$$

Согласно (14.3-2) известно, что p_2 имеет ошибку

$$\frac{h^3}{3} y'''(\theta_1)$$

и по (14.3-4) c_2 имеет ошибку

$$-\frac{h^3}{12} y'''(\theta_2).$$

Если y''' приблизительно постоянна в нашем интервале, то

$$p - c = \frac{5h^3}{12} y'''. \quad (14.4-2)$$

Таким образом, ошибка в корректирующем значении примерно равна $-\frac{1}{5}(p - c)$. Следовательно, надо добавить $\frac{1}{5}(p - c) = -0,002$, чтобы получить окончательное значение

$$y = c + \frac{1}{5}(p - c) = 0,422.$$

Теперь можно повторять цикл, пока не получим $x = 1$ (см. таблицу 14.4-1).

Таблица 14.4-1

Численное решение уравнения $y' = y^2 + 1$ ($y(0) = 0$)

x	y	y'	p	p'	c	$p - c$	$\frac{p - c}{5}$
0	0	1					
0,2	0,203	1,0412					
0,4	0,422	1,178	0,41648	1,173	0,424	-0,008	-0,002
0,6	0,683	1,466	0,674	1,455	0,685	-0,011	-0,002
0,8	1,027	2,047	1,008	2,016	1,031	-0,023	-0,004
1,0	1,546		1,502	3,256	1,557	-0,053	-0,011
Корректир. значение Ошибки	1,557 0,011						

В качестве проверки находим, что верное значение $y(1) = 1,557$, тогда как найденное нами $y(1) = 1,546$. Абсолютная ошибка 0,011 дает относительную ошибку

$$\frac{0,011}{1,557} \times 100 = 0,7 \text{ процента,}$$

которая довольно хороша, если учитывать разбиение и использование логарифмической линейки.

Упражнения

14.4-1. Решить численно уравнение $y' = y$, $y(0) = 1$, взяв $h = 0,2$ для $0 \leq x \leq 1$.

14.4-2. Решить численно уравнение $y' = y^2 - \sin^2 x$, $y(0) = 0$, взяв $h = \frac{\pi}{6}$ для $0 \leq x \leq 2\pi$.

§ 14.5. Устойчивость метода простого прогноза

Иногда предполагают, что формулу прогноза (14.3-1)

$$y_{n+1} = y_{n-1} + 2hy'_n \quad (14.5-1)$$

можно использовать без дальнейшей корректировки. Исследуем устойчивость такого метода.

Пусть $z(x)$ — точное решение уравнения

$$z' = f(x, z), \quad (14.5-2)$$

а y_n — вычисленное решение при $x = x_n$. Если значение y_n подставить в дифференциальное уравнение, то оно удовлетворит ему, так как вычисленное значение y'_n было найдено из уравнения. Таким образом, пренебрегая ошибками округления, имеем

$$y'_n = f(x, y_n). \quad (14.5-3)$$

Вычитая (14.5-3) из (14.5-2), получаем

$$\varepsilon'_n = z'_n - y'_n = f(x, z_n) - f(x, y_n) = \frac{\partial f(x, \theta)}{\partial z} \varepsilon_n = A \varepsilon_n, \quad (14.5-4)$$

где $A = \frac{\partial f}{\partial z}$, а θ лежит между y_n и z_n .

Когда мы подставляем верное решение в разностное уравнение прогноза, то получается ошибка $e(x)$, так как верное решение, вообще говоря, не удовлетворяет этому разностному уравнению, т. е.

$$z_{n+1} = z_{n-1} + 2hz'_n + e_n \quad (14.5-5)$$

Теперь, вычитая уравнение (14.5-1) из (14.5-5), получаем

$$\varepsilon_{n+1} = \varepsilon_{n-1} + 2h\varepsilon'_n + e_n \quad (14.5-6)$$

Мы занимаемся оценкой локального роста ошибки ε_n . Поэтому можно принять, что e_n и $\frac{\partial f}{\partial y} = A$ постоянны. Если же они изменяются

не медленно, значит шаг интегрирования слишком велик. При этих предположениях можно подставить (14.5-4) в (14.5-6); получается

$$\varepsilon_{n+1} = \varepsilon_{n-1} + 2hA\varepsilon_n + e_n$$

или

$$\varepsilon_{n+1} - 2Ah\varepsilon_n - \varepsilon_{n-1} = e_n \quad (14.5-7)$$

что является линейным разностным уравнением с постоянными коэффициентами.

Характеристическое уравнение этого разностного уравнения есть

$$\rho^2 - 2Ah\rho - 1 = 0. \quad (14.5-8)$$

Следовательно, его решение

$$\varepsilon_n = C_1(\rho_1)^n + C_2(\rho_2)^n, \quad (14.5-9)$$

где

$$\rho_1 = Ah + \sqrt{A^2h^2 + 1}, \quad \rho_2 = Ah - \sqrt{A^2h^2 + 1}. \quad (14.5-10)$$

Если $A = \frac{\partial f}{\partial y} > 0$, то $\rho_1 > 1$, а если $A < 0$, то $|\rho_2| > 1$. В любом случае один из двух членов ρ_1^n или ρ_2^n становится большим при $n \rightarrow \infty$. В этом случае говорят, что метод неустойчив.

Мы видим, что при сделанных предположениях повторное применение прогноза (14.5-1) является неустойчивым методом. (См., однако, § 14.6.)

Упражнение 14.5-1. Попробуйте численно решить уравнение $y' = -y$, $y(0) = 1$, $h = 0,1$ для $0 \leq x \leq 1$, используя лишь прогноз, и сравните результаты с теорией.

§ 14.6. Устойчивость коррекции

Предположение, что коррекция повторяется до тех пор, пока не прекратится изменение y , может быть исследовано тем же способом, каким мы пользовались для прогноза. Формула коррекции

$$y_{n+1} = y_n + \frac{h}{2} (y'_{n+1} + y'_n) \quad (14.6-1)$$

приводит аналогичным путем к характеристическому уравнению

$$\begin{aligned} \left(1 - \frac{Ah}{2}\right)\rho - \left(1 + \frac{Ah}{2}\right) &= 0, \\ \rho &= \frac{1 + \frac{Ah}{2}}{1 - \frac{Ah}{2}} = 1 + Ah + \frac{(Ah)^2}{2} + \frac{(Ah)^3}{4} + \dots \end{aligned} \quad (14.6-2)$$

(при условии, что $\left|\frac{Ah}{2}\right| < 1$).

При рассмотрении итерационного процесса, применяемого для коррекции, исследуем сначала $\frac{Ah}{2}$. Пусть c_i — значения y_{n+1} в i -й итерации. Вычислим поправку

$$\begin{aligned} c_{i+1} - c_i &= \frac{h}{2} [f(x, c_i) - f(x, c_{i-1})] = \\ &= \frac{h}{2} \frac{\partial f}{\partial c_i} (c_i - c_{i-1}) \approx \frac{Ah}{2} (c_i - c_{i-1}). \end{aligned}$$

Очевидно, процесс будет сходиться, если коэффициент сходимости

$$\left| \frac{Ah}{2} \right| < 1. \quad (14.6-3)$$

Возвращаясь к (14.6-2), замечаем, что ρ дается первыми тремя членами разложения e^{Ah} плюс незначительная ошибка в четвертом члене.

В простом уравнении

$$y' = Ay \quad (A > 0)$$

имеем

$$\frac{\partial f}{\partial y} = A$$

и решение в точке x_n есть

$$y_n = Ce^{Ah_n},$$

что совпадает с (14.6-2) (если пренебречь членами с h^3).

Таким образом, видно, что мы заблуждаемся, считая критерием неустойчивости выполнение неравенства $|\rho_i| \approx |e^{Ah}| < 1$. Очевидно, если решение растет как геометрическая прогрессия, ошибки, которые растут с той же самой скоростью или меньше, не обязательно будут губить решение. С другой стороны, если $A < 0$, то ошибки должны убывать по крайней мере так же быстро, как $e^{(Ah)n}$, иначе они будут преобладать в вычисленном решении. Тот факт, что ошибки ограничены, не помогает в случае убывающего решения. Вообще говоря, важным фактором при оценке ошибки является число кривых — решений, которые мы пересекаем.

Все сказанное наводит на мысль, что от решений дифференциальных уравнений нужно требовать относительной устойчивости, которая определяется как скорость роста ошибки по отношению к росту решения. Когда относительная устойчивость по модулю меньше единицы, шум вследствие изолированного округления или другой ошибки будет расти не быстрее, чем решение.

С точки зрения этого нового критерия нам надо пересмотреть выводы § 14.5 об использовании простого прогноза. Если предполагать,

что $|Ah| < 1$, то характеристические корни ρ_i уравнения (14.5-10) могут быть представлены в виде

$$\rho_1 = Ah + \sqrt{1 + (Ah)^2} = Ah + 1 + \frac{(Ah)^2}{2} - \frac{1}{8}(Ah)^4 + \dots,$$

$$\rho_2 = Ah - \sqrt{1 + (Ah)^2} = Ah - \left[1 + \frac{(Ah)^2}{2} - \frac{1}{8}(Ah)^4 + \dots \right],$$

или

$$\begin{aligned} \rho_1 &= 1 + Ah + \frac{(Ah)^2}{2} + \dots \approx e^{Ah}, \\ \rho_2 &= -\left[1 - Ah + \frac{(Ah)^2}{2} + \dots \right] \approx -e^{-Ah}, \end{aligned} \quad (14.6-4)$$

а само решение растет как e^{Ah} . Отсюда видно, что если $A > 0$, то можно надеяться, что повторное использование прогноза даст разумные результаты, но если $A < 0$, то $(\rho_2)^n$ будет расти по величине и колебаться, в то время как верное решение убывает. Следовательно, метод повторного применения прогноза относительно нестабилен только, когда $A < 0$ (см. упражнение 14.5-1).

Возвращаясь к итерационному процессу корректировки, мы видим, что характеристическое уравнение имеет лишь один корень (уравнение (14.6-2)) и решение ведет себя так же, как и ошибка; таким образом, итерационный метод коррекции относительно стабилен.

Упражнение 14.6-1. Показать, что при использовании простого прогноза результат численного интегрирования

$$y' = y, \quad y(0) = 1, \quad h = 0,2 \quad \text{для} \quad 0 \leq x \leq 1$$

согласуется с теорией.

§ 14.7. Несколько общих замечаний

Мы дали четыре примера анализа устойчивости: первый относится к неопределенным интегралам, второй — к особому интегралу со сверткой, последние два — к дифференциальным уравнениям. Заметим, что устойчивость для интегралов и для дифференциальных уравнений — не одно и то же; для дифференциальных уравнений существуют пути обратной связи от производных, чего нет в случае интегралов. Необходимость анализа устойчивости возникает всякий раз, когда старые значения используются для вычисления новых, и таким образом получается обратная связь ошибок. Не существует простого правила, относящегося ко всем случаям, но ясно, что устойчивость тесно связана с обратной связью. На последнюю тему написано много книг. Было показано также, что для дифференциальных уравнений идея устойчивости ведет по ложному пути и относительная устойчивость является более разумным критерием.

Анализ устойчивости, который был проведен на простом примере предыдущего параграфа, был неполным. Мы проанализировали прогноз и коррекцию отдельно, пренебрегая эффектами их взаимодействия и игнорируя заключительную часть метода — уравнение (14.4-2). Для исследования, конечно, можно написать сложную систему дифференциальных уравнений и построить соответствующее характеристическое уравнение, но результаты будут очень запутанными. Мы предпочтем сделать несколько замечаний о том, как взаимодействуют прогноз и коррекция. Неустойчивость прогноза имеет очень маленькое влияние на относительную устойчивость результата после примененной один раз коррекции. Заключительный этап также не вызывает больших затруднений по сравнению с исследованием относительной устойчивости коррекции.

Ту же идею, которая была использована, чтобы уничтожить ошибку скорректированного значения, можно использовать, чтобы подправить прогнозируемое значение и, таким образом, уменьшить ошибку в оценке прогнозируемой производной, входящей в окончательное значение. Чтобы сделать это, следует прибавить к прогнозируемому значению $-\frac{4}{5}(p_n - c_n)$ последнего цикла. Если существует тенденция к колебанию, это только подчеркнет ее. С другой стороны, нет сомнения, что если неустойчивость и округление незначительны, такой член хорошо влияет на точность всего вычисления.

Упражнение 14.7-1. Разработать формулы прогноза, поправки, коррекции и окончательного значения для интегрирования системы.

§ 14.8. Системы уравнений

Дифференциальное уравнение n -го порядка

$$y^{(n)} = f(x, y, y', \dots, y^{(n-1)})$$

может быть сведено к системе n уравнений первого порядка простым изменением в обозначениях. Напишем

$$\begin{aligned} y &= y_0, & y' &= y_1, & y'' &= y_1' = y_2, \\ y''' &= y_2' = y_3, & \dots, & & y^{(n-1)} &= y_{n-1} \end{aligned}$$

и мы имеем эквивалентную систему:

$$\begin{aligned} y_0' &= y_1, \\ y_1' &= y_2, \\ y_2' &= y_3, \\ &\dots \dots \dots \\ y_{n-1}' &= f(x, y_0, y_1, y_2, \dots, y_{n-1}). \end{aligned}$$

Чтобы применить изложенный метод интегрирования к системе, мы просто применим операторы одновременно ко всем уравнениям. Связи между уравнениями проявляются только при вычислении производных. Другими словами, уравнения обрабатываются отдельно, но параллельно.

Метод поля направлений обычно применяется лишь к одному уравнению первого порядка, но в случае системы двух уравнений первого порядка, в которых правая часть не зависит от переменной x :

$$y' = f(y, z), \quad z' = g(y, z),$$

деление одного уравнения на другое дает уравнение первого порядка без x . Поле направлений для этого уравнения дает y в зависимости от z , а иногда достаточно показать, как ведет себя y в зависимости от z , не определяя x , для которого точка (y, z) встретилась на кривой.

ГЛАВА 15

ОБЩАЯ ТЕОРИЯ МЕТОДОВ ПРОГНОЗА И КОРРЕКЦИИ

§ 15.1. Введение

В главах 13 и 14 были введены основные понятия и указаны некоторые технические приемы реализации методов прогноза и коррекции для численного интегрирования обыкновенных дифференциальных уравнений. Теперь мы используем все это, чтобы развить общую теорию методов прогноза и коррекции. Эта теория включает методы Милна и Адамса—Башфорта как частные случаи. Таким образом, мы можем сравнить эти широко применяемые методы с другими в рамках единой теории.

Методы прогноза и коррекции для интегрирования обыкновенных дифференциальных уравнений широко используются благодаря следующим преимуществам:

1. Разность между прогнозированным и скорректированным значениями дает оценку ошибок, сделанных на каждом шагу и, следовательно, может быть использована для контроля величины шага, применяемого при интегрировании.
2. На каждом шагу надо вычислять производную один или два раза по сравнению с четырьмя для метода Рунге—Кутты (гл. 16), что для систем высокого порядка может значительно экономить время на вычисления.
3. Легко ловить различные ошибки.

Некоторый недостаток состоит в том, что методы сложны для программирования и не являются «самоначинающимися». Этот вопрос будет обсужден в гл. 16.

При использовании методов прогноза и коррекции для интегрирования обыкновенных дифференциальных уравнений возникают разного рода трудности. Вот их главные источники:

1. Ошибки от отбрасывания членов, которые возникают при аппроксимации производных.

2. Распространение ошибок — неустойчивость, которая возникает при решении аппроксимирующих разностных уравнений, не соответствующих решениям исходных дифференциальных уравнений.

3. Усиление ошибок округления из-за комбинаций коэффициентов в конечноразностных формулах.

В этой главе мы достаточно подробно исследуем широкий класс формул и покажем, что выбор конкретной формулы есть компромисс между весьма противоречивыми желаниями. Так как при интегрировании обыкновенных дифференциальных уравнений формула коррекции играет наиболее важную роль, естественно исследовать коррекцию сначала и притом более внимательно.

После исследования коррекции мы займемся изучением прогноза. Хорошо ознакомившись как с прогнозом, так и с коррекцией, мы сможем затем исследовать задачу выбора шага и оценки точности решения при интегрировании обыкновенных дифференциальных уравнений. Наконец, мы сделаем выводы и обсудим некоторые экспериментальные результаты.

Исследованные методы имеют ошибку от отбрасывания членов порядка h^3 ; опыт работы показывает, что ими пользуются наиболее широко. В гл. 14 был разработан метод прогноза и коррекции порядка h^3 . Предоставляем читателю самостоятельно разработать промежуточную теорию методов порядка h^4 .

Метод третьего порядка гл. 14 не имеет свободных параметров; методы четвертого порядка, изложенные в упражнениях, вводят два новых коэффициента, один из которых используется, чтобы повысить порядок ошибки от отбрасывания членов, а другой — для устойчивости; методы пятого порядка имеют еще два коэффициента, из которых снова один используется для отбрасывания членов и один для устойчивости, и т. д. для методов более высокого порядка.

В большинстве случаев ошибка от отбрасывания членов при конечной аппроксимации производных дифференциального уравнения или системы дифференциальных уравнений на каждом шагу значительно превосходит ошибки округления. Однако при построении специализированных вычислительных машин стремятся сделать длину чисел и, следовательно, основной уровень точности, как можно меньшими. В таких ситуациях ошибка от отбрасывания членов может быть примерно той же величины, что и ошибка округления, и последнюю

нельзя игнорировать при выборе методов интегрирования для данной задачи. Кроме того, в наши дни все более часто случается, что в некоторых задачах, таких как вычисление траекторий Луны и планет, счет с восьми- и даже десятизначной точностью может давать значительные ошибки округления, и тогда некоторые из описанных здесь методов становятся полезными.

§ 15.2. Ошибка от отбрасывания членов

Самая общая формула линейной коррекции, использующая информацию о функции и первой производной в последних трех точках решения и оценку производной в точке, которая должна быть вычислена, имеет вид

$$y_{n+1} = a_0 y_n + a_1 y_{n-1} + a_2 y_{n-2} + h(b_{-1} y'_{n+1} + b_0 y'_n + b_1 y'_{n-1} + b_2 y'_{n-2}) + E_8 \frac{h^3 y^{(3)}(\theta)}{5!}. \quad (15.2-1)$$

Можно использовать и другие линейные формы, но эта включает большое число общих методов и, кроме того, ею мы уже пользовались в гл. 13 (уравнение (13.3-1)), когда исследовали неопределенные интегралы. Однако в отличие от гл. 13 теперь мы имеем обратную связь не только через значения старого решения, но и через члены с производными; следовательно, изучение устойчивости будет здесь более сложным.

Выписанная формула предполагается точной для многочленов до четвертой степени, т. е. для $y = 1, x, x^2, x^3, x^4$. Используя эти пять условий и взяв a_1 и a_2 в качестве параметров, получаем (см. уравнение (13.3-2))

$$\left. \begin{aligned} a_0 &= 1 - a_1 - a_2, & b_0 &= \frac{1}{24}(19 + 13a_1 + 8a_2), \\ a_1 &= a_1, & b_1 &= \frac{1}{24}(-5 + 13a_1 + 32a_2), \\ a_2 &= a_2, & b_2 &= \frac{1}{24}(1 - a_1 + 8a_2), \\ b_{-1} &= \frac{1}{24}(9 - a_1), & E_8 &= \frac{1}{6}(-19 + 11a_1 - 8a_2). \end{aligned} \right\} \quad (15.2-2)$$

Пять коэффициентов из семи используются, чтобы уменьшить ошибку от отбрасывания членов. При помощи двух других мы, вместо того чтобы еще дальше уменьшать ошибку от отбрасывания членов, постараемся уменьшить ошибки второго и третьего типов. Как и в гл. 14, разность прогнозируемого и скорректированного значений будет использована как для того, чтобы дать оценку ошибки, так и для того, чтобы «уничтожить» главный член ошибки.

Обычный метод нахождения остаточного члена основан на использовании ряда Тейлора (см. (11.3-2) и (13.4-1))

$$y(x) = y(A) + (x-A)y'(A) + \frac{(x-A)^2}{2!}y''(A) + \dots \\ \dots + \frac{(x-A)^n}{n!}y^{(n)}(A) + \frac{1}{n!} \int_A^x y^{(n+1)}(s)(x-s)^n ds$$

с $A = -2h$ и $n = 4$. Следуя гл. 13, получаем

$$R(y) = \int_{-2h}^h y^{(5)}(s) G(s) ds,$$

где $G(s)$ — функция влияния:

$$G(s) = (h-s)_+^4 - a_0(-s)_+^4 - a_1(-h-s)_+^4 - a_2(-2h-s)_+^4 - \\ - 4h[b_{-1}(h-s)_+^3 + b_0(-s)_+^3 + b_1(-h-s)_+^3 + b_2(-2h-s)_+^3] \quad (15.2-3)$$

и (см. (11.3-4))

$$(a-s)_+^k = \begin{cases} (a-s)^k, & \text{если } a-s \geq 0 \\ 0, & \text{если } a-s < 0. \end{cases} \quad (k > 0),$$

Если $G(s)$ имеет постоянный знак, то существует такое θ ($-2h < \theta < h$), что ошибка может быть записана в виде

$$y^{(5)}(\theta) \int_{-2h}^h G(s) ds. \quad (15.2-4)$$

Если $G(s)$ меняет знак, то найдутся функции $f(s)$, для которых ошибка не может быть представлена в виде (15.2-4).

Исследование § 13.4 дает показанную на рис. 13.4-1 и 13.4-2 область постоянства знака $G(s)$. Она-то и представляет для нас основной интерес. На тех же рисунках для нее показаны линии равной ошибки от отбрасывания членов.

§ 15.3. Устойчивость

Следуя § 14.5, допустим, что $z = z(x)$ есть решение

$$z' = f(x, z), \quad (15.3-1)$$

$$z_{n+1} = a_0 z_n + a_1 z_{n-1} + a_2 z_{n-2} + \\ + h(b_{-1} z'_{n+1} + b_0 z'_n + b_1 z'_{n-1} + b_2 z'_{n-2}) + e_n, \quad (15.3-2)$$

а $y = y_n$ — вычисленное решение.

$$y'_n = f(x_n, y_n), \quad (15.3-3)$$

$$y_{n+1} = a_0 y_n + a_1 y_{n-1} + a_2 y_{n-2} + \\ + h(b_{-1} y'_{n+1} + b_0 y'_n + b_1 y'_{n-1} + b_2 y'_{n-2}). \quad (15.3-4)$$

Положим $\varepsilon_n = z_n - y_n$ и, вычитая соответственно равенства (15.3-2) и (15.3-4), получим

$$\varepsilon_{n+1} = a_0 \varepsilon_n + a_1 \varepsilon_{n-1} + a_2 \varepsilon_{n-2} + \\ + h(b_{-1} \varepsilon'_{n+1} + b_0 \varepsilon'_n + b_1 \varepsilon'_{n-1} + b_2 \varepsilon'_{n-2}) + e_n, \quad (15.3-5)$$

тогда как (15.3-1) и (15.3-3), если использовать теорему о среднем значении, дают

$$\varepsilon'_n = f(x_n, z_n) - f(x_n, y_n) = \frac{\partial f(x_n, \theta)}{\partial z} \varepsilon_n, \quad (15.3-6)$$

где θ — обычное среднее значение.

При изучении роста ошибки ε_n разумно допустить, что e_n и $\frac{\partial f}{\partial z} = A$ обе постоянны; на практике они медленно изменяются от шага к шагу. Подставив (15.3-6) в (15.3-5), получаем

$$(1 - b_{-1}Ah) \varepsilon_{n+1} = (a_0 + b_0Ah) \varepsilon_n + (a_1 + b_1Ah) \varepsilon_{n-1} + \\ + (a_2 + b_2Ah) \varepsilon_{n-2} + e_n, \quad (15.3-7)$$

т. е. линейное разностное уравнение с постоянными коэффициентами.

В уравнении (15.3-7) не встречается порознь ни h , ни A , но всюду только величина Ah . Хотя Ah может меняться в широких пределах, надо разумно оценить значения, которые можно ожидать. Чтобы хоть как-то представить себе их, рассмотрим простые уравнения

$$y' = \pm Ay, \quad y(0) = 1.$$

Ошибку от отбрасывания членов на каждом шагу примем равной $\frac{h^5 A^5}{40} \left(\frac{1}{40} \right)$ — это типичное значение для $\frac{E_5}{5!}$; см. таблицу 13.7-1). Мы приходим к таблице 15.3-1.

Таблица 15.3-1

Ошибка от отбрасывания членов как функция от $|Ah|$

$ Ah $	0,1	0,2	0,3	0,4	0,5
$\frac{h^5 y^{(5)}}{40}$	$2,50 \times 10^{-7}$	$8,00 \times 10^{-6}$	$6,07 \times 10^{-5}$	$2,56 \times 10^{-4}$	$7,88 \times 10^{-4}$

Таким образом, чтобы ошибка от отбрасывания членов была достаточно мала, необходимо, чтобы $|Ah| \leq \frac{5}{10}$ и для точных решений $|Ah| \leq \frac{2}{10}$.

Конечно, другие уравнения будут иметь другие формулы для более высоких производных (возможно, производные будут расти быстрее, чем простые степени), и вообще нельзя сказать ничего конкретного относительно того, какой будет $|Ah|$ по сравнению с $\frac{E_5 h^5 y^{(5)}}{5!}$. Тем не менее мы примем для большинства интересующих нас уравнений, что $|Ah| < 0,4$.

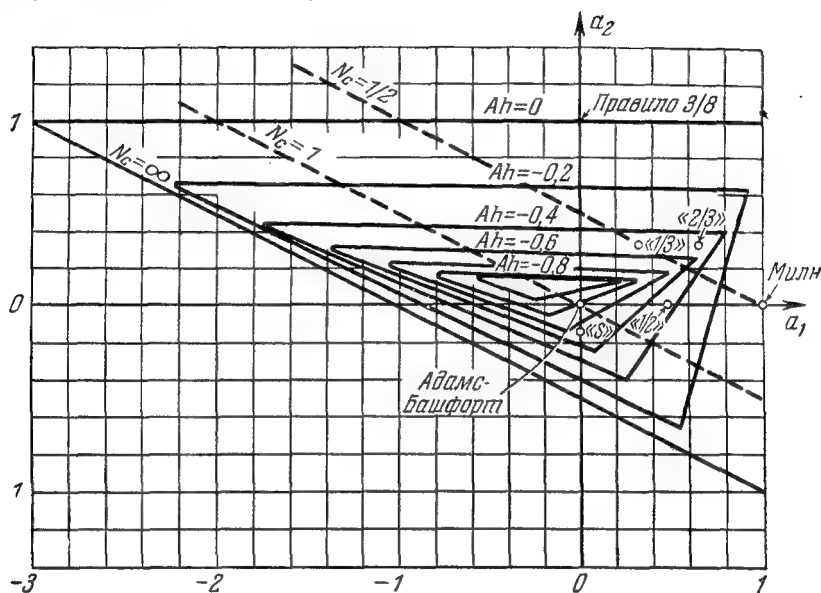


Рис. 15.3-1. Область устойчивости.

На рис. 15.3-1 показана область устойчивости в случае $Ah = 0$. Когда $Ah \neq 0$, однородное разностное уравнение, соответствующее (15.3-7), приводит к характеристическому уравнению

$$Ah = \frac{\rho^3 - a_0 \rho^2 - a_1 \rho - a_2}{b_{-1} \rho^3 + b_0 \rho^2 + b_1 \rho + b_2}. \quad (15.3-8)$$

Это уравнение более сложное, чем соответствующее уравнение в гл. 13 из-за дополнительной обратной связи от членов с производными. Решение (ср. с (13.5-6)) есть

$$\epsilon_n = C_1 (\rho_1)^n + C_2 (\rho_2)^n + C_3 (\rho_3)^n. \quad (15.3-9)$$

Решение первоначального дифференциального уравнения (15.3-1) имеет тенденцию расти как

$$y_n = C e^{(Ah)n},$$

и вообще то, с чем мы будем иметь дело, — это не абсолютный рост ошибки, а ее рост относительно локального роста решения. Таким образом, мы добиваемся не устойчивости, когда $|p_t| < 1$, а относительной устойчивости, когда

$$|pe^{-Ah}| \leq 1. \quad (15.3-10)$$

Один из корней, который по-прежнему будет обозначаться p_3 , ведет себя так

$$p_3 = 1 + Ah + \frac{(Ah)^2}{2} + \frac{(Ah)^3}{3!} + \frac{(Ah)^4}{4!} + k \frac{(Ah)^5}{5!} \approx e^{Ah} \quad (15.3-11)$$

для маленьких $|Ah|$. Таким образом,

$$p_3 e^{-Ah} \approx 1$$

и, следовательно, лишь p_1 и p_2 могут привести к относительной неустойчивости.

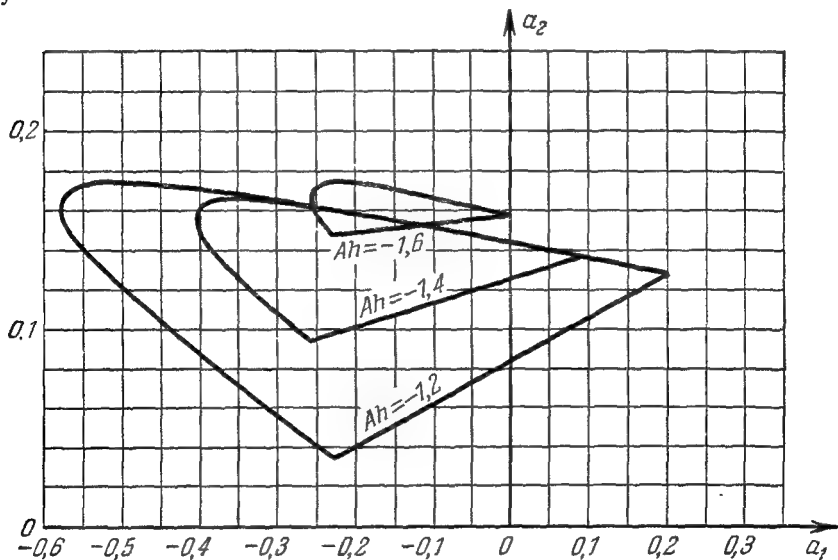


Рис. 15.3-2. Область устойчивости.

Треугольник, вдоль которого при $Ah = 0$ корни принимают значения

$$|pe^{-Ah}| = |p| = 1,$$

переходит при Ah , не равном нулю, в фигуру несколько более сложную. Для $p = -e^{Ah}$ получается линейное уравнение

$$\alpha a_2 + \beta a_1 + \gamma = 0, \quad (15.3-12)$$

где α , β и γ — комбинации Ah , e^{Ah} , e^{2Ah} и e^{3Ah} . Но другие две стороны треугольника становятся гиперболами. На рис. 15.3-1 и 15.3-2

показана для отрицательных значений Ah одна ветвь гиперболы до пересечения с прямой (15.3-12). Значения внутри другой ветви должны быть устойчивы, а в промежутке между ветвями должны быть неустойчивые значения, так что они, кроме, возможно, специальных случаев, не могут быть полезны. Рис. 15.3-1 и 15.3-2 показывают области устойчивости для корней p_1 и p_2 . Для больших значений $|Ah|$ корень p_3 далек от значения e^{Ah} и ошибкой от отбрасывания членов нельзя пренебрегать.

Упражнения

15.3-1. Начертить прямую (15.3-12) для $Ah > 0$ и исследовать область устойчивости.

15.3-2. Используя упражнение 13.7-1, разработать соответствующую теорию для дифференциальных уравнений. Найти область устойчивости в зависимости от a_1 и Ah .

§ 15.4. Помехи округления

Теория распространения помех округления для дифференциального уравнения очень похожа на соответствующую теорию для неопределенных интегралов, изложенную в § 13.6. Для $Ah = 0$ прямая (13.6-3) $a_2 = -\frac{a_1}{2}$ разделяет область роста и область убывания ошибки округления. Внутри области устойчивости коэффициент усиления шума дается формулой (13.6-2)

$$N_c = \frac{1}{1 + a_1 + 2a_2}.$$

Когда $Ah \neq 0$, граница между двумя областями более сложна, но здесь не стоит этим заниматься; просто нужно помнить, что лучше быть выше линии, чем ниже ее.

Чтобы не был слишком большим некоррелированный шум, хорошо сохранять

$$N_a = (a_0^2 + a_1^2 + a_2^2)^{1/2}$$

достаточно малым.

§ 15.5. Прогноз по трем точкам

В соответствии с формулой коррекции (15.2-1) хотелось бы иметь прогноз, использующий информацию о последних трех точках, а именно:

$$y_{n+1} = A_0 y_n + A_1 y_{n-1} + A_2 y_{n-2} + \\ + h(B_0 y'_n + B_1 y'_{n-1} + B_2 y'_{n-2}) + \bar{E}_5 \frac{h^5 y^{(5)}}{5!} + \dots \quad (15.5-1)$$

Коэффициенты можно найти или обычным путем, или положив в уравнении (15.2-2) $b_{-1} = 0$. Результаты

$$\left. \begin{aligned} A_0 &= -8 - A_2, & B_0 &= \frac{17 + A_2}{3}, \\ A_1 &= 9, & B_1 &= \frac{14 + 4A_2}{3}, \\ & & B_2 &= \frac{-1 + A_2}{3}, \\ A_2 &= A_2, & \bar{E}_3 &= \frac{40 - 4A_2}{3}. \end{aligned} \right\} \quad (15.5-2)$$

Большое значение A_1 , равное девяти, означает, что даже в самом лучшем случае получится большое усиление шума

$$N_a = (A_0^2 + A_1^2 + A_2^2)^{1/2} > 9$$

в прогнозе.

Для того чтобы в прогнозе по трем точкам избежать большого усиления шума, можно сделать формулу точной только до x^3 включительно, оставив остаточный член порядка h^4 . Это приводит к семейству прогнозов:

$$\left. \begin{aligned} A_0 &= 1 - A_1 - A_2, & B_1 &= \frac{-16 + 8A_1 + 16A_2}{12}, \\ A_1 &= A_1, & B_2 &= \frac{5 - A_1 + 4A_2}{12}, \\ A_2 &= A_2, & & \\ B_0 &= \frac{23 + 5A_1 + 4A_2}{12}, & \bar{E}_4 &= 9 - A_1. \end{aligned} \right\} \quad (15.5-3)$$

Чтобы воспользоваться разностью n -го прогноза и коррекции для оценки ошибки $(n+1)$ -го прогноза, следует определить p_{n+1} из

$$m_{n+1} = p_{n+1} - (p_n - c_n) = c_n + (p_{n+1} - p_n).$$

Такая схема иногда бывает полезной, хотя может требовать очень большого количества вычислений, особенно, когда требуется точное решение дифференциальных уравнений. Мы не будем обсуждать ее дальше; ясно, что она имеет остаточный член пятого порядка.

§ 15.6. Прогнозы типа Милна

Если мы хотим иметь семь параметров, чтобы использовать их в поисках подходящего прогноза, то можно прибавить или одно более старое значение функции, а именно y_n , или одно старое значение производной, а именно y'_n . Первое предложение включает прогноз Милна, и мы назовем его прогнозом «типа Милна», а второе содержит прогноз Адамса—Башфорта и называется соответственно.

Попробуем сначала найти формулу вида

$$y_{n+1} = A_0 y_n + A_1 y_{n-1} + A_2 y_{n-2} + A_3 y_{n-3} + \\ + h(B_0 y'_n + B_1 y'_{n-1} + B_2 y'_{n-2}) + \bar{E}_5 \frac{h^5 y^{(5)}}{5!} + \dots \quad (15.6-1)$$

Обычный метод дает

$$\left. \begin{aligned} A_0 &= -8 - A_2 + 8A_3, & B_0 &= \frac{17 + A_2 - 9A_3}{3}, \\ A_1 &= 9 - 9A_3, & B_1 &= \frac{14 + 4A_2 - 18A_3}{3}, \\ A_2 &= A_2, & B_2 &= \frac{-1 + A_2 + 9A_3}{3}, \\ A_3 &= A_3, & \bar{E}_5 &= \frac{40 - 4A_2 + 72A_3}{3}. \end{aligned} \right\} \quad (15.6-2)$$

Чтобы остаточный член этой формулы имел указанный вид, необходимо, чтобы функция влияния

$$G(s) = (h-s)_+^4 - A_0(-s)_+^4 - A_1(-h-s)_+^4 - A_2(-2h-s)_+^4 - \\ - A_3(-3h-s)_+^4 - 4h[B_0(-s)_+^3 + B_1(-h-s)_+^3 + B_2(-2h-s)_+^3]$$

была постоянного знака для $h \geq s \geq -3h$. В плоскости (A_2, A_3) линия, вдоль которой $G(s) = 0$, движется снизу к оси $A_3 = 0$, когда s меняется от h к $-2h$, и остается на $A_3 = 0$ для $-2h \geq s \geq -3h$. Таким образом, чтобы остаточный член имел нужный вид, ограничимся случаем $A_3 \geq 0$.

Если попробовать уменьшить ошибку от отбрасывания членов, выбрав $A_3 = 0$, то найдем, что $A_1 = 9$, а это достаточно неприятно. С другой стороны, если попробовать минимизировать усиление шума

$$N_a = (A_0^2 + A_1^2 + A_2^2 + A_3^2)^{1/2},$$

то получим

$$A_0 = -\frac{4}{114}, \quad A_1 = \frac{9}{114}, \quad A_2 = -\frac{4}{114}, \quad A_3 = \frac{113}{114}.$$

Это очень близко к прогнозу Милна

$$y_{n+1} = y_{n-3} + \frac{4h}{3}(2y'_n - y'_{n-1} + 2y'_{n-2}) + \frac{14}{45}h^5 y^{(5)}(\theta)$$

с $A_0 = A_1 = A_2 = 0$ и $A_3 = 1$. Выигрыш кажется несостоящим.

Таким образом, исследуя этот метод, мы приходим к прогнозу Милна. Как было замечено в § 14.4, его можно объединить в схему гл. 14 с любой подходящей коррекцией. Конечно, возможен и другой выбор параметров A_2 и A_3 .

Упражнение 15.6-1. Исследовать прогнозы вида

$$y_{n+1} = A_0 y_n + A_1 y_{n-1} + A_2 y_{n-2} + h(B_0 y'_n + B_1 y'_{n-1}) + \bar{E}_4 \frac{h^4 y^{(IV)}(\theta)}{4!},$$

используя A_2 как параметр.

$$\text{О т в е т: } A_0 = -4 - 5A_2, \quad A_1 = 5 + 4A_2, \quad A_2 = A_2,$$

$$B_0 = 4 + 2A_2, \quad B_1 = 2 + 4A_2, \quad \bar{E}_4 = 4 - 4A_2.$$

Если $A_2 = 1$, то $\bar{E}_4 = 0$ и $\bar{E}_5 = 12$.

§ 15.7. Прогнозы типа Адамса—Башфорта

Если вместо того, чтобы воспользоваться дополнительным значением функции, возьмем дополнительное значение производной y'_{n-3} , то получим формулу

$$y_{n+1} = A_0 y_n + A_1 y_{n-1} + A_2 y_{n-2} + \\ + h(B_0 y'_n + B_1 y'_{n-1} + B_2 y'_{n-2} + B_3 y'_{n-3}) + \bar{E}_5 \frac{h^5 y^{(5)}}{5!}, \quad (15.7-1)$$

коэффициенты которой равны.

$$\left. \begin{aligned} A_0 &= 1 - A_1 - A_2, & B_1 &= \frac{-59 + 19A_1 + 32A_2}{24}, \\ A_1 &= A_1, & B_2 &= \frac{37 - 5A_1 + 8A_2}{24}, \\ A_2 &= A_2, & B_3 &= \frac{-9 + A_1}{24}, \\ B_0 &= \frac{55 + 9A_1 + 8A_2}{24}, & \bar{E}_5 &= \frac{251 - 19A_1 - 8A_2}{6}. \end{aligned} \right\} \quad (15.7-2)$$

В этом случае функция влияния $G(s)$ имеет нули в плоскости (A_1, A_2) вдоль линий, первая из которых — прямая $A_1 = 9$ — для $s = -2h$ поднимается и приобретает отрицательные тангенсы угла наклона с увеличением s , не накладывая существенных ограничений на коэффициенты.

Ряд случаев, включая прогноз Адамса—Башфорта, приведен в таблице 15.7-1. Значение коэффициента усиления шума N_c для $h \rightarrow 0$ здесь то же самое, что и для коррекции, и его можно не повторять.

Упражнение 15.7-1. Исследовать прогнозы вида

$$y_{n+1} = A_0 y_n + A_1 y_{n-1} + h(B_0 y'_n + B_1 y'_{n-1} + B_2 y'_{n-2}) + \bar{E}_4 \frac{h^4 y^{(IV)}(\theta)}{4!},$$

используя A_1 как параметр.

$$\text{О т в е т: } A_0 = 1 - A_1; \quad B_0 = \frac{23 + 5A_1}{12}; \quad B_2 = \frac{5 - A_1}{12};$$

$$A_1 = A_1; \quad B_1 = \frac{-16 + 8A_1}{12}; \quad \bar{E}_4 = 9 - A_1.$$

Таблица 15.7-1

Прогнозы типа Адамса — Башфорта

[Значения $\bar{E}_s - E_s$ вычислены для случая $a_i = A_i$ ($i=0, 1, 2$)]

	Адамс — Башфорт	Высокая точность	Правило «3/8»	«1/3»	«1/2»	«2, 3»
A_0	1	0	0	$\frac{1}{3}$	$\frac{1}{2}$	0
A_1	0	1	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{2}{3}$
A_2	0	0	1	$\frac{1}{3}$		$\frac{1}{3}$
B_0	$\frac{55}{24}$	$\frac{8}{3}$	$\frac{21}{8}$	$\frac{94}{36}$	$\frac{110}{48}$	$\frac{191}{72}$
B_1	$\frac{59}{24}$	$-\frac{5}{3}$	$-\frac{9}{8}$	$-\frac{63}{36}$	$-\frac{99}{48}$	$-\frac{107}{72}$
B_2	$\frac{37}{24}$	$\frac{1}{3}$	$\frac{15}{8}$	$\frac{57}{36}$	$\frac{69}{48}$	$\frac{108}{72}$
B_3	$\frac{9}{24}$	$-\frac{1}{3}$	$-\frac{3}{8}$	$-\frac{13}{36}$	$-\frac{17}{48}$	$-\frac{25}{72}$
\bar{E}_s	$\frac{251}{6}$	$\frac{110}{3}$	$\frac{243}{6}$	$\frac{121}{3}$	$\frac{161}{4}$	$\frac{707}{18}$
$\bar{E}_s - E_s$	$\frac{270}{6}$	$\frac{210}{6}$	$\frac{270}{6}$	$\frac{260}{6}$	$\frac{255}{6}$	$\frac{250}{6}$

§ 15.8. Общие замечания о выборе метода

При построении конкретных формул прогноза и коррекции приходится уравнивать три до некоторой степени несовместимые цели:

1. Уменьшение ошибки от отбрасывания членов.
2. Большой запас устойчивости.
3. Защита от неполадок с округлением.

Лучше всего начать с выбора коррекции, так как она играет более важную роль. Рис. 13.5-1 показывает область, из которой можно выбрать коррекцию. Ясно, что коррекция Милна ($a_1=1$, $a_2=0$) примерно так же точна, как наилучшая, какая может быть найдена. Рис. 15.3-1 показывает меру устойчивости. Если $A = \frac{\partial f}{\partial y}$ положительно на всем интервале интегрирования, то метод Милна хорош. Но если $A = \frac{\partial f}{\partial y}$ может быть отрицательным, то коррекция

Милна неустойчива, и нужно выбирать компромиссное решение между двумя неприятностями: неустойчивостью и ошибкой от отбрасывания членов, измеряемыми соответственно величинами

$$\frac{\partial f}{\partial y} h \equiv Ah \quad \text{и} \quad \frac{E_5 h^5 y^{(5)}(0)}{5!}.$$

Между этими двумя выражениями нет определенного соотношения, хотя ясно, что оба имеют тенденцию быть или большими, или маленькими по величине одновременно.

Мы положим $Ah = -0,4$, считая, что при расчетах с умеренной точностью такой выбор для большинства уравнений оградит нас от беспокойств относительно неустойчивости. Сравнение рис. 13.5-1 и 15.3-1 показывает, что вдоль прямой части контура $Ah = -0,4$ ошибка от отбрасывания членов почти одинакова.

Неполадки с округлением наводят на мысль подвинуться вправо вверх и выбрать для коррекции точку $a_1 = \frac{2}{3}$, $a_2 = \frac{1}{3}$. В некотором смысле это на $\frac{2}{3}$ метод Милна и на $\frac{1}{3}$ «правило трех восьмых». Милн и Рейнольдс*) предлагают использовать метод Милна, периодически применяя правило трех восьмых, но предложенное смешение этих методов на каждом шагу более удовлетворительно обеспечивает устойчивость.

Хотя, взяв $Ah = -0,4$, мы надеялись вообще избавиться от неустойчивости, следует заметить, что при решении уравнений с большим отрицательным $\frac{\partial f}{\partial y}$ нужно выбирать более устойчивый метод. Из рис. 15.3-1 видно, что точка ($a_1 = -0,20$, $a_2 = 0,15$) дает очень устойчивый метод, но зато он имеет достаточно большую ($E_5 = -\frac{56}{15}$) ошибку от отбрасывания членов.

Упражнение 15.8-1. В упражнении 15.3-2 исследовать выбор корректирующей формулы четвертого порядка.

§ 15.9. Выбор прогноза

При выборе формулы прогноза существует сильный соблазн использовать в этой формуле для A_i те же значения, какие будут приняты для a_i в формуле коррекции. Тогда при вычислении траекторий движения Луны или планет можно вычислять линейную комбинацию значений y (положений) программой, считающей с двойной, а выражения для y' — с обычной точностью и притом достичь большого увеличения точности.

*) J. Assoc. Computing Machinery, vol. 6, pp. 196—203, апрель 1959.

Такое соответствие коэффициентов A_i и a_i исключает, конечно, прогнозы типа Милна, имеющие несколько большую точность, чем методы Адамса. Здесь мы исследуем только те случаи, в которых подобный выбор A_i и a_i возможен.

Разность между прогнозированным и скорректированным значениями y_{n+1} оценивает пятую производную. При равенстве A_i и a_i члены с y уничтожаются и мы имеем

$$p_{n+1} - c_{n+1} = h [B_0 y'_n + B_1 y'_{n-1} + B_2 y'_{n-2} + B_3 y'_{n-3} - (b_{-1} p'_{n+1} + b_0 y'_n + b_1 y'_{n-1} + b_2 y'_{n-2})], \quad (15.9-1)$$

что вообще может быть вычислено совершенно точно.

Упражнение 15.9-1. Исследовать соответствующую теорию методов четвертого порядка.

§ 15.10. Некоторые формулы

Можно использовать тот факт, что $p_n - c_n$ обычно более или менее постоянна от шага к шагу, для уничтожения ошибки в прогнозе, равной

$$\frac{\bar{E} h^5 y^{(5)}}{5!}.$$

Чтобы уничтожить эту ошибку, естественно использовать величину

$$-\frac{\bar{E}_5}{\bar{E}_5 - E_5} (p_n - c_n) = -\frac{\bar{E}_5 h^5 y^{(5)}}{5!}.$$

(На первом шагу, когда отсутствует $p_n - c_n$, мы положим $p_n - c_n = 0$.) Тем же способом можно попытаться исправить коррекцию; поправку естественно взять равной

$$\frac{E_5}{\bar{E}_5 - E_5} (p_{n+1} - c_{n+1}) \approx \frac{E_5 h^5 y^{(5)}}{5!}.$$

Применим все это, например, к так называемому «правилу двух третьих». Формулы в этом случае имеют следующий вид:

Прогноз:

$$p_{n+1} = \frac{2y_{n-1} + y_{n-2}}{3} + \frac{h}{72} (191y'_n - 107y'_{n-1} + 109y'_{n-2} - 25y'_{n-3}) + \frac{707}{2160} h^5 y^{(5)}.$$

Изменение:

$$m_{n+1} = p_{n+1} - \frac{707}{750} (p_n - c_n).$$

Коррекция:

$$c_{n+1} = \frac{2y_{n-1} + y_{n-2}}{3} + \\ + \frac{h}{72} (25m'_{n+1} + 91y'_n + 43y'_{n-1} + 9y'_{n-2}) - \frac{43}{2160} h^5 y^{(5)}.$$

Окончательное значение:

$$y_{n+1} = c_{n+1} + \frac{43}{750} (p_{n+1} - c_{n+1}).$$

Здесь получается точный ответ, если $y^{(5)}(x)$ постоянна, так что этот метод можно назвать методом шестого порядка.

Следует заметить, что если y_{n+1} близко к m_{n+1} , то можно не вычислять производную y'_{n+1} , а использовать значение m'_{n+1} . Если $\varepsilon = y'_{n+1} - m'_{n+1}$, то, предполагая, что $|Ah| = 0,4$, в следующем прогнозированном значении получим ошибку, равную

$$\varepsilon h \frac{191}{72} \left| \frac{\partial f}{\partial y} \right| \leq \varepsilon \cdot 2,65 \cdot |A| \cdot h \approx 1,06\varepsilon.$$

Другим случаем, заслуживающим внимания, является правило одной второй (которое есть среднее арифметическое методов Милна и Адамса — Башфорта):

$$p_{n+1} = \frac{y_n + y_{n-1}}{2} + \frac{h}{48} (119y'_n - 99y'_{n-1} + 69y'_{n-2} - 17y'_{n-3}) + \frac{161}{480} h^5 y^{(5)},$$

$$m_{n+1} = p_{n+1} - \frac{161}{170} (p_n - c_n)$$

$$c_{n+1} = \frac{y_n + y_{n-1}}{2} + \frac{h}{48} (17m'_{n+1} + 51y'_n + 3y'_{n-1} + y'_{n-2}) - \frac{9}{480} h^5 y^{(5)},$$

$$y_{n+1} = c_{n+1} + \frac{9}{170} (p_{n+1} - c_{n+1}).$$

Этот метод имеет несколько меньшую ошибку от отбрасывания членов, чем «правило двух третьих». Однако $N_c = \frac{2}{3}$ и $N_a = 0,7071$ по сравнению с $N_c = \frac{3}{5}$ и $N_a = 0,75$ для «правила двух третьих».

Упражнение 15.10-1. Построить соответствующую теорию для методов четвертого порядка и получить конкретные формулы.

§ 15.11. Выбор шага и оценка точности

Кроме формул интегрирования для каждого шага, необходимо иметь формулы для деления пополам и удвоения шага и критерий, указывающий на необходимость изменения шага. Нам нужно также получить несколько первых значений решения, для чего мы используем здесь метод Рунге — Кутты (см. гл. 16).

Существуют две причины для деления пополам (или уменьшения каким-либо другим способом) величины шага интегрирования: первая — большая ошибка от отбрасывания членов, вторая — неустойчивость. Защита от неустойчивости во время вычисления зависит от умения оценить $A = \frac{\partial f}{\partial y}$. Если мы оценим производные и для m_{n+1} и для y_{n+1} , получим

$$f(x_{n+1}, m_{n+1}) - f(x_{n+1}, y_{n+1}) \approx \frac{\partial f(x_{n+1}, y_{n+1})}{\partial y} (m_{n+1} - y_{n+1}) \approx A(m_{n+1} - y_{n+1}).$$

Периодическая оценка A при помощи двух вычислений производной на одном шаге необходима для того, чтобы обезопасить себя от неприятностей с неустойчивостью, когда нет других способов оценки величины A .

Если обнаружено, что A слишком велико для обычной величины шага h , то можно либо уменьшить h , либо обратиться к более устойчивой формуле. Если при этом ошибка от отбрасывания членов мала, то скорее всего хорошо заменить формулу на более устойчивую, хотя и с большей ошибкой от отбрасывания членов. С другой стороны, если обе ошибки близки к допуску, то следует уменьшить h .

Мы неоднократно показывали, что ошибка от отбрасывания членов измеряется величиной

$$p_n - c_n \approx \frac{(\bar{E}_5 - E_5) h^5 y^{(5)}}{5!}.$$

Но уже когда используются формулы, предложенные в предыдущем параграфе, это не так. Действительно, там же было замечено, что если бы $y^{(5)}$ была константой, то ошибка от отбрасывания членов была бы равна нулю. Истинная ошибка измеряется не величиной $p_n - c_n$, а изменением $p_n - c_n$ от шага к шагу. Кажется соблазнительным взять в качестве меры ошибки величину

$$(p_{n+1} - c_{n+1}) - (p_n - c_n)$$

и в зависимости от ее значения решать вопрос о том, нужно ли делить шаг пополам. К сожалению, эта величина очень сильно зависит от местного шума в вычислении (исключая, конечно, случаи, когда с большим числом знаков ведутся вычисления очень низкой точности). Поэтому использовать ее в качестве критерия необходимости уменьшения шага рискованно. Таким образом, мы оказываемся в очень неудачном положении, зная, что ошибка шага от отбрасывания членов много меньше, чем величина $p_n - c_n$, но не зная насколько.

Оценка скорости изменения $p_n - c_n$ скажем, с помощью такой разности

$$(p_n - c_n) - (p_{n-10} - c_{n-10}),$$

быть может, и полезна, но здесь использоваться не будет.

Чтобы уменьшить шаг вдвое, нужно иметь значения $y_{n-1/2}$ и $y_{n-3/2}$. Имея y_n , y_{n-1} , y_{n-2} и их производные, мы можем вычислить $y_{n-1/2}$ по ним, сделав формулу точной для 1, x , ..., x^3 (что делает ошибку зависящей от $y^{(6)}$). Такую формулу легко получить, используя упражнение 10.5-1:

$$y_{n-1/2} = \frac{1}{128} [45y_n + 72y_{n-1} + 11y_{n-2} + h(-9y'_n + 36y'_{n-1} + 3y'_{n-2})].$$

Чтобы получить $y_{n-3/2}$, можно перевернуть формулу:

$$y_{n-3/2} = \frac{1}{128} [11y_n + 72y_{n-1} + 45y_{n-2} - h(3y'_n + 36y'_{n-1} - 9y'_{n-2})].$$

В тех случаях, когда шаг нужно удвоить, можно поступать одним из следующих способов:

1. Хранить дополнительное старое значение и использовать его при удвоении шага.

2. Начать сначала.

3. Использовать для первых двух шагов специальные формулы.

Первый метод, кроме всего прочего, нехорош тем, что требует специальной заботы о том, чтобы не было двух удвоений шага подряд и чтобы старое значение существовало (не вылезало за начало).

В некоторых программах вопрос об увеличении или уменьшении шага вдвое решается путем сравнения величины $p_{n+1} - c_{n+1}$ с двумя константами c_d и c_h . Когда требуется разделить шаг пополам, только что вычисленные значения в $n+1$ точке отбрасываются, интервал делится пополам интерполяцией и шаги повторяются опять. Однако, если требуется деление пополам на первом шаге прогноза и коррекции, то все начало должно быть повторено.

Следует позаботиться о том, чтобы сохранить соотношение $c_h > > 100c_d$. В противном случае существует риск, что из-за случайных колебаний машина может уменьшить или увеличить последующие шаги.

Может быть принята иная точка зрения: если машинные ошибки вообще вероятны, то внезапный скачок $|p_n - c_n|$, возможно, происходит скорее вследствие сбоя машины, чем из-за слишком большого шага. Метод, при котором в случае деления шага последнее посчитанное значение отбрасывается, исключает машинную ошибку за счет деления пополам и последующего удвоения. Чтобы следить за внезапными скачками, можно использовать специальный тест, и если такой скачок действительно произошел, шаг может быть просто повторен.

§ 15.12. Экспериментальная проверка

Простейший подход к задаче выбора метода для численного интегрирования некоторого дифференциального уравнения состоит в использовании вычислительной машины для решения разными методами ряда специально подобранных случаев (выбираемых обычно так, чтобы они имели известные решения). Хотя такой способ проверки иногда необходим, он редко бывает достаточным.

Чтобы проверить только что изложенную теорию, было испытано на трех уравнениях большое количество методов.

Уравнение	Начальное условие	Аналитическое решение	
$y' = y$	$y(0) = 1$	$y = e^x$	(15.12-1)

$y' = -y$	$y(0) = 1$	$y = e^{-x}$	(15.12-2)
-----------	------------	--------------	-----------

$y' = -2xy^2$	$y(0) = 1$	$y = \frac{1}{1+x^2}$	(15.12-3)
---------------	------------	-----------------------	-----------

Уравнения решались для $0 \leq x \leq 10$ при различной величине шага h . Были выбраны именно эти уравнения из-за того, что первые два имеют, по существу, «локально линейный режим» общего уравнения и ясно обнаруживают неполадки из-за неустойчивости. Третье имеет решение, которое часто является трудным для аппроксимации многочленами.

Использовались лишь методы, в которых формулы прогноза и коррекции имеют одни и те же коэффициенты для значений функции y ; кроме того, был также испытан классический метод Милна.

Так как во всех примерах вычисление производной ведется по очень простой формуле, в половине испытаний для воспроизведения шума округления был добавлен равномерный случайный шум.

Результаты, которые слишком объемны, чтобы приводить их здесь, могут быть кратко изложены следующим образом.

1. Были проверены условия устойчивости. Использование модификации, основанной на $p_n - c_n$ подчеркивает колебания вследствие неустойчивости, а использование $p_{n-1} - c_{n-1}$ помогает предупредить эту тенденцию к неустойчивости.

2. Для умеренно точных решений использование модификации прогноза и окончательного значения уменьшало ошибку примерно в 10 раз (при условии, что соблюдалась устойчивость), так что

$$\frac{1}{10} |p_{n+1} - c_{n+1}| \approx |m_{n+1} - y_{n+1}|$$

Всякий раз, когда величина шага, а следовательно, и ошибки оказывались слишком маленькими, эти два шага не помогали, так как округление преобладало над маленькими эффектами коррекции.

3. Для уравнения $y' = y$ правило «трех восьмых», несмотря на его большую ошибку от отбрасывания членов, было очень хорошим.

4. Если исключить неустойчивые методы, то в среднем методы Адамса — Башфорта, «двух третьих» и «одной второй», по-видимому, не различаются существенно; возможно лишь некоторое преимущество у первого.

Эти результаты не следует принимать как окончательные, и дальнейшее изучение специальных ситуаций, возможно, раскроет много больше. Область численного интегрирования дифференциальных уравнений является очень сложной, и полное изучение ее не относится к вводному курсу. Поэтому мы оставим тему экспериментальной проверки, не создав, надеемся, впечатления, что она не является важной.

ГЛАВА 16

СПЕЦИАЛЬНЫЕ МЕТОДЫ ИНТЕГРИРОВАНИЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

§ 16.1. Введение и общее описание

Методы прогноза и коррекции предыдущей главы требуют специальных приемов для нахождения достаточного числа начальных значений, чтобы начать прогноз. Вероятно, самыми широко используемыми начальными методами являются методы Рунге — Кутта. Хотя эти методы можно применять и на каждом шагу, эффективность в этом случае, вероятно, будет низкой. Типичный метод Рунге — Кутта требует четырех вычислений производных для каждого шага вперед, тогда как методы прогноза и коррекции требуют одного или двух на шаг. В то же время оба метода имеют почти одинаковую ошибку от отбрасывания членов. Методы Рунге — Кутта не содержат в себе никакой оценки точности решения на каждом шагу; следовательно, невозможно догадаться, когда нужно уменьшить или удвоить величину шага. Высказывалась идея использования двух интегрирований с разной величиной шага, но это, очевидно, так увеличивает машинное время, что предложение не следует принимать всерьез.

Когда система уравнений должна решаться многократно, стоит исследовать уравнения и посмотреть, нельзя ли использовать их особенности, чтобы уменьшить количество требуемых вычислений. Так, например, линейные дифференциальные уравнения второго порядка (с переменными коэффициентами) имеют достаточно специфическую структуру. Поэтому применение специальных методов существенно уменьшает количество вычислений по сравнению с тем, когда уравнение второго порядка записывается в канонической форме, т. е. в виде двух уравнений первого порядка, и затем решается по общей схеме.

Изобретено великое множество специальных методов. Здесь мы можем рассмотреть только некоторые из них; поэтому мы просто покажем, как можно вывести пару специальных методов. Построения иллюстрируют те же общие идеи, какие были использованы в предыдущих главах, и еще раз показывают силу общих методов.

§ 16.2. Методы Рунге — Кутта

Существует много вариантов метода Рунге — Кутта; наиболее широко используется следующий: дано

$$y' = f(x, y), \quad y(x_n) = y_n \quad (16.2-1)$$

вычисляется последовательно

$$\left. \begin{aligned} k_1 &= hf(x_n, y_n), \\ k_2 &= hf\left(x_n + \frac{h}{2}, y_n + \frac{k_1}{2}\right), \\ k_3 &= hf\left(x_n + \frac{h}{2}, y_n + \frac{k_2}{2}\right), \\ k_4 &= hf(x_n + h, y_n + k_3), \\ y_{n+1} &= y_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4). \end{aligned} \right\} \quad (16.2-2)$$

Этот процесс можно представить геометрически. В точке (x_n, y_n) вычисляется тангенс угла наклона (k_1/h) ; используя его, мы идем на половину шага вперед и смотрим тангенс угла наклона здесь. Используя новый тангенс угла наклона (k_2/h) , мы опять начинаем из (x_n, y_n) , идем вперед на половину шага и опять берем пробу тангенса угла наклона. Взяв этот последний тангенс угла наклона (k_3/h) , мы опять начинаем из (x_n, y_n) , но делаем теперь полный шаг вперед, где смотрим тангенс угла наклона (k_4/h) . Четыре тангенса углов наклона усредняем с весами $1/6, 2/6, 2/6, 1/6$ и, беря этот средний тангенс угла наклона, делаем окончательный шаг от (x_n, y_n) к (x_{n+1}, y_{n+1}) .

Если $f(x, y)$ не зависит от y , то усреднение можно производить по формуле Симпсона. Этот метод имеет остаточный член, пропорциональный h^5 .

Существуют варианты метода, касающиеся того, где следует брать узловые точки и, следовательно, какие веса придавать им на различных шагах. Существуют также методы более низкого порядка, использующие меньше узловых точек на шаг.

Очевидно, в этом методе пропадает зря вся предыдущая информация и каждый полный шаг делается заново, а следовательно, этот метод едва ли является таким же эффективным, как методы, которые используют старую информацию. Очевидно также, что не существует

проверки того, является ли шаг слишком маленьким или слишком большим, и хотя, возможно, изучение k_i может дать ключ к этому, его обычно не применяют.

Кажется, метод получения этих уравнений не используется в других местах численного анализа, и поэтому его не стоит давать здесь *). Общий смысл построения их в том, что функции $f(x, y)$, которые находятся справа, все расписываются в ряды по степеням h и соответствующие производные приравняются, чтобы устранить более низкие степени h . В результате уравнения (16.2-2) являются точными для многочленов степени 4 (или ниже).

§ 16.3. Методы для уравнения второго порядка, когда отсутствует y'

Можно исследовать много частных случаев, но в этом параграфе мы ограничимся общей ситуацией; рассмотрим дифференциальное уравнение второго порядка с отсутствующей первой производной. Часто, когда производная присутствует, простое преобразование удаляет ее. Для исследования этой проблемы применим те же методы, какие использовались в гл. 15.

Начнем с коррекции при интегрировании дифференциального уравнения

$$y'' = f(x, y), \quad (16.3-1)$$

которую представим в виде

$$y_{n+1} = a_0 y_n + a_1 y_{n-1} + a_2 y_{n-2} + h^2 (b_{-1} y_{n+1}'' + b_0 y_n'' + b_1 y_{n-1}'' + b_2 y_{n-2}''). \quad (16.3-2)$$

Метод последней главы может навести на мысль экономить два параметра, но простая проверка показывает, что использование a_2 в качестве единственного параметра позволит нам надежно решить проблему устойчивости. Таким образом, мы можем сделать метод точным для 1, x, \dots, x^5 . Результаты приведены в таблице 16.3-1.

Метод в последней колонке имеет нулевой остаточный член; следовательно, он имеет более высокую точность, чем остальные. Но кратный характеристический корень означает, что распространяющаяся ошибка ϵ_n будет вести себя как

$$\epsilon_n = C_1 + C_2 n + C_3 n^2,$$

что наводит на грустные размышления. В других случаях имеем

$$\epsilon_n = C_1 + C_2 n + C_3 (a_2)^n.$$

*) См. [20].

Т а б л и ц а 16.3-1

	$a_2 = -1$	$a_2 = -1/2$	$a_2 = 0$	$a_2 = 1/2$	$a_2 = 1$
$a_0 = 2 + a_2$	1	3/2	2	5/2	3
$a_1 = -(1 + 2a_2)$	1	0	-1	-2	-3
$a_2 = a_2$	-1	-1/2	0	1/2	1
$b_{-1} = 1/12$	1/12	1/12	1/12	1/12	1/12
$b_0 = (10 - a_2) : 12$	11/12	21/24	10/12	19/24	9/12
$b_1 = (1 - 10a_2) : 12$	11/12	12/24	1/12	-8/24	-9/12
$b_2 = -a_2 : 12$	1/12	1/24	0	-1/24	-1/12
$E_6 = -3 + 3a_2$	-6	-9/2	-3	-3/2	0
$\rho_1, \rho_2, \rho_3 = 1, 1, a_2$	1, 1, -1	1, 1, -1/2	1, 1, 0	1, 1, 1/2	1, 1, 1

Кратный корень $\rho = 1$ является неизбежным, если формула делается точной для 1 и x , а линейный рост ошибки происходит от ошибки в y' (которая не появляется явно), распространяясь линейно с n .

Преимущество двух нулевых коэффициентов при выборе (см. [26], формула (44.1)) $a_2 = 0$ делает формулу очень привлекательной

$$y_{n+1} = 2y_n - y_{n-1} + \frac{h^2}{12} (y''_{n+1} + 10y''_n + y''_{n-1}) - \frac{h^6 y^{(6)}}{240}. \quad (16.3-3)$$

Это можно также записать в виде

$$y_{n+1} = 2y_n - y_{n-1} + h^2 \left(y''_n + \frac{\Delta^2 y''_{n-1}}{12} \right) - \frac{h^6 y^{(6)}}{240}. \quad (16.3-4)$$

По симметрии легко проверить, что $G(s) \neq 0$ для $-h < s < h$.

Удобный прогноз для этой коррекции такой (приводим его здесь без вывода):

$$y_{n+1} = 2y_{n-1} - y_{n-3} + \frac{4h^2}{3} (y''_n + y''_{n-1} + y''_{n-2}) + \frac{16}{240} h^6 y^{(6)}, \quad (16.3-5)$$

или

$$y_{n+1} = 2y_{n-1} - y_{n-3} + 4h^2 \left(y''_{n-1} + \frac{\Delta^2 y''_{n-2}}{3} \right) + \frac{16}{240} h^6 y^{(6)}. \quad (16.3-6)$$

«Прогноз минус коррекция» дает

$$\frac{17}{240} h^6 y^{(6)},$$

так что ошибка при коррекции приблизительно равна

$$-\frac{1}{17} (p_{n+1} - c_{n+1})$$

в то время как ошибка в прогнозе

$$\frac{16}{17}(p_n - c_n).$$

Этот результат может быть использован, чтобы улучшить точность.

Характеристические корни прогноза суть 1, -1 , i , $-i$ и не создают трудностей.

Упражнения

16.3-1. Вывести равенство (16.3-5).

16.3-2. Показать, что справедливо равенство (16.3-5), исследовав функцию G .

§ 16.4. Линейные уравнения

Линейные дифференциальные уравнения встречаются очень часто, и, несмотря на обширную теорию, связанную с ними, иногда приходится решать их численными методами. Свойство линейности делает их более легкими для решения, чем общие уравнения. В качестве примера рассмотрим уравнение

$$y'' = f(x)y + g(x), \quad (16.4-1)$$

которое является частным случаем уравнения, рассмотренного в § 16.3 (так как y' не содержится в нем явно). Используя коррекцию в виде (16.3-4), имеем

$$\Delta^2 y_{n-1} - \frac{h^2}{12} \Delta^2 y_{n-1}'' = h^2 y_n'' - \frac{h^6 y^{(6)}}{240}.$$

Подставляя (16.4-1) вместо y'' , находим

$$\Delta^2 y_{n-1} - \frac{h^2}{12} \Delta^2 [f(x_{n-1})y_{n-1} + g(x_{n-1})] = h^2 [f(x_n)y_n + g(x_n)],$$

$$\Delta^2 \left\{ \left[1 - \frac{h^2}{12} f(x_{n-1}) \right] y_{n-1} \right\} = h^2 f(x_n) y_n + h^2 \left(g_n + \frac{1}{12} \Delta^2 g_{n-1} \right).$$

Теперь напомним

$$\left[1 - \frac{h^2}{12} f(x_n) \right] y_n = Y_n$$

так что

$$\Delta^2 Y_{n-1} = h^2 \left[f(x_n) y_n + g_n + \frac{1}{12} \Delta^2 g_{n-1} \right].$$

Таким образом, окончательно

$$Y_{n+1} = 2Y_n - Y_{n-1} + h^2 \left[f(x_n) y_n + g_n + \frac{1}{12} \Delta^2 g_{n-1} \right], \quad (16.4-2)$$

где

$$y_n = \frac{Y_n}{1 - \frac{h^2}{12} f(x_n)}. \quad (16.4-3)$$

Этот метод часто приписывают Нумерову *); он очень полезен и эффективен. Очевидно, что подобные преобразования, сделанные с целью исключить прогноз, могут быть произведены всякий раз, когда уравнение линейное, и мы не будем рассматривать все возможные случаи.

Упражнения

16.4-1. Вычислить решение

$$y'' + xy = 0, \quad y(0) = 1, \quad y'(0) = 0$$

для $0 \leq x \leq 1$, $h = \frac{2}{10}$, используя метод Нумерова.

16.4-2. Показать, что не требуется прогноза для линейного уравнения

$$y' = P(x)y + Q(x).$$

§ 16.5. Метод, который использует значения y , y' и y''

Мы отмечали несколько раз, что когда функция вычислена, часто бывает недорого (в смысле машинного времени) вычислить и производную функции. Итак, пусть задано уравнение

$$y' = f(x, y); \quad (16.5-1)$$

тогда

$$y'' = \frac{\partial f}{\partial y} y' + \frac{\partial f}{\partial x}. \quad (16.5-2)$$

Упражнение 13.8-1 дает формулу ($b_{-1} = 1/2$)

$$y_{n+1} = y_n + \frac{h}{2} (y'_{n+1} + y'_n) + \frac{h^2}{12} (-y''_{n+1} + y''_n) + \frac{h^5}{720} y^{(5)}(\theta). \quad (16.5-3)$$

По-видимому, это — превосходный выбор для формулы коррекции. Так как встречается лишь член с y_n , то не существует посторонних корней характеристического уравнения, порождающих неустойчивость. Ясно, что ошибка от отбрасывания членов очень мала и округление также находится под контролем. Мы должны поэтому найти прогноз в общем виде

$$y_{n+1} = A_0 y_n + A_1 y_{n-1} + h(B_0 y'_n + B_1 y'_{n-1}) + h^2(C_0 y''_n + C_1 y''_{n-1}).$$

Обычный процесс нахождения коэффициентов, делающих прогноз точным для 1, x , x^2 , x^3 , x^4 , дает результаты, приведенные в таблице 16.5-1. В частности, при $A_1 = 1$ получаются очень привлекательные коэффициенты, хотя $A_1 = 0$ дает несколько меньшую ошибку.

*) См. [12].

Т а б л и ц а 16.5-1

	$A_1 = 0$	$A_1 = \frac{1}{2}$	$A_1 = 1$		$A_1 = 0$	$A_1 = \frac{1}{2}$	$A_1 = 1$
$A_0 = 1 - A_1$	1	$\frac{1}{2}$	0	$C_0 = (17 - A_1):12$	$\frac{17}{12}$	$\frac{33}{24}$	$\frac{4}{3}$
$A_1 = A_1$	0	$\frac{1}{2}$	1	$C_1 = (7 + A_1):12$	$\frac{7}{12}$	$\frac{15}{24}$	$\frac{2}{3}$
$B_0 = (-1 + A_1):2$	$-\frac{1}{2}$	$-\frac{1}{4}$	0	$E_0 = (31 + A_1):6$	$\frac{31}{6}$	$\frac{21}{4}$	$\frac{16}{3}$
$B_1 = (3 + A_1):2$	$\frac{3}{2}$	$\frac{7}{4}$	2				

Упражнения

16.5-1. Разработать детали метода прогноза и коррекции для случая $A_1 = 1$, включая модификации прогноза и окончательных значений.

16.5-2. Исследовать функцию влияния для случая $A_1 = 1$.

§ 16.6. Случай, когда решение трудно аппроксимировать многочленом

До сих пор всюду предполагалось, что рассматриваемые функции хорошо приближаются многочленами на интервале разумной длины. Единственное исключение было сделано в § 12.1, где рассматривались интегралы, представимые в виде

$$\int_a^b f(x) dx = \int_a^b K(x) g(x) dx.$$

Мы заметили там, что если известны моменты

$$m_k = \int_a^b K(x) x^k dx$$

и функция $g(x)$ хорошо аппроксимируется многочленом, то интеграл может быть легко вычислен, независимо от того, хорошо ли приближается многочленом функция $f(x)$. Таким образом аппроксимировалась лишь часть подынтегральной функции.

Такие ситуации возникают и при решении дифференциальных уравнений; часто решения, которые ищутся, очень плохо приближаются многочленами разумной степени при разумной величине шага. И если известно, как ведет себя какая-либо функция, входящая в решение, то можно попытаться использовать это, аппроксимируя не решение, а лишь часть его.

Для примера предположим, что решается уравнение

$$ay'' + by' + cy = f(x), \quad y(x_0) = y_0, \quad y'(x_0) = y'_0, \quad (16.6-1)$$

где a , b и c — константы и $f(x)$ легко аппроксимировать; скажем,

$$f(x) \approx p_i + q_i(x - x_{i-1}) \quad \text{для} \quad x_{i-1} \leq x \leq x_i.$$

Таким образом, $f(x)$ аппроксимируется последовательностью прямолинейных отрезков. Непосредственной подстановкой легко найти частное решение (16.6-1)

$$y(x) = F_i(x) = P_i + Q_i(x - x_{i-1}).$$

Решение однородного дифференциального уравнения приводит к характеристическому уравнению

$$am^2 + bm + c = 0$$

с корнями

$$m_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}, \quad m_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}.$$

Следовательно, решение (16.6-1) в i -м интервале ($x_{i-1} \leq x \leq x_i$) начинается точкой

$$y(x_{i-1}) = y_{i-1}, \quad y'(x_{i-1}) = y'_{i-1}$$

и дается формулой

$$y(x) = C_1 e^{m_1(x-x_{i-1})} + C_2 e^{m_2(x-x_{i-1})} + F_i(x).$$

Чтобы определить C_1 и C_2 используем начальные условия для интервала

$$y_{i-1} = C_1 + C_2 + F_i(x_{i-1}), \quad y'_{i-1} = m_1 C_1 + m_2 C_2 + F'_i(x_{i-1}).$$

Теперь вычислим начальные условия для следующего интервала:

$$y(x_i) = y_i, \quad y'(x_i) = y'_i$$

и мы готовы к следующему шагу.

Если одно или оба значения m_i отрицательные и большие, то решение плохо аппроксимируется многочленом и, написав решение в виде суммы многочлена и экспоненциальной части, мы избежим многих неприятностей. Использованный метод никоим образом не зависит от того, что для аппроксимации $f(x)$ были взяты линейные многочлены, и, вообще говоря, многочлены более высоких порядков были бы более экономичны в смысле машинного времени. Метод не связан и с выбранным нами уравнением второго порядка, порядок уравнения мог бы быть любым.

Исследуем теперь другой класс задач, в которых аппроксимируем многочленом не окончательный ответ, а только некоторую часть его.

Большие отрицательные характеристические корни типичны для многих задач, возникающих в контрольных системах, при получении скорости изменения сигнала, в задачах распределения пространственного заряда, в химии горения. Простейший вид уравнений такого рода

$$y' = f(x, y), \quad \text{где} \quad \frac{\partial f}{\partial y} \approx A \ll 0.$$

Один из подходов к этому уравнению основан на предположении, что производная $\frac{\partial f}{\partial y}$, хотя большая и отрицательная, изменяется медленно. Мы пишем ($\frac{\partial f}{\partial y}$ приблизительно равно $A = \text{const}$)

$$y' = f(x, y) + Ay - Ay, \quad y' - Ay = f(x, y) - Ay \equiv F(x, y),$$

где, конечно,

$$\frac{\partial F}{\partial y} = \frac{\partial f}{\partial y} - A \approx 0.$$

Решая уравнение при помощи обычного интегрирующего множителя e^{-Ax} , имеем

$$y(x) = C_1 e^{A(x-x_0)} + e^{A(x-x_0)} \int_{x_0}^x e^{-A(\theta-x_0)} F(\theta, y(\theta)) d\theta.$$

Теперь мы сталкиваемся с задачей получения формул для интегралов вида

$$I(h) = \int_0^h e^{-A\theta} g(\theta) d\theta,$$

которую легко решить, следуя уже развитым методам. Прогноз не может использовать значение $g(h)$, но может использовать $g(0)$, $g(-h)$, а также $I(0)$, $I(-h)$, ... тогда как в корректирующую формулу может входить прогнозируемое значение $g(h)$. Детали оставляем читателю. Здесь предполагается лишь, что разность

$$F(x, y) = f(x, y) - Ay \quad [y - y(x)]$$

может быть аппроксимирована многочленом с использованием довольно редкой группы узловых точек, хотя во многих случаях это вовсе не так.

В первом из вышеприведенных примеров неаппроксимируемая часть входила слагаемым, во втором она была множителем. Но в обоих случаях мы отказались от прямой аппроксимации решения многочленами в пользу аппроксимации некоторой его части.

Упражнение 16.6-1. Предположим, что $f(x)$ может быть аппроксимирована квадратным трехчленом в интервале $0 \leq x \leq k \cdot 2\pi$, и рассмотрим уравнение $y'' + y = f(x)$. Построить теорию для численного интегрирования этого уравнения в данном интервале. Показать, как распространить ее последовательно на интервалы $2k\pi \leq x \leq 4k\pi$, $4k\pi \leq x \leq 6k\pi$ и т. д.

§ 16.7. Краевые задачи

До сих пор рассматривались задачи с начальными условиями. Часто бывают даны условия на решение в двух точках. Например, может быть дано уравнение

$$y'' = f(x, y), \quad y(0) = 0, \quad y(1) = 0$$

и требуется знать $y = y(x)$ для $0 \leq x \leq 1$.

Эта ситуация может быть сведена к предыдущему случаю методом проб и ошибок. Мы начинаем с уравнения

$$y'' = f(x, y), \quad y(0) = 0, \quad y'(0) = \lambda$$

и пробуем найти два таких значения λ_1 и λ_2 , чтобы для

$$\lambda_1: y(1) < 0, \quad \lambda_2: y(1) > 0.$$

Так как для большинства практических задач $y(1)$ есть непрерывная функция λ , то можно, даже используя грубый метод деления отрезка пополам, за 10 проб уменьшить длину отрезка $|\lambda_1 - \lambda_2|$ в 2^{10} , т. е. более чем в 1000 раз. Впрочем, для нахождения значения λ , при котором $y(1) = 0$, легко найти гораздо более эффективные методы.

Однако, вместо того чтобы сводить краевую задачу к задаче с заданными начальными условиями, нередко выгоднее решать ее непосредственно. На этот счет имеется обширная теория ([11]), и мы кратко рассмотрим лишь один пример, чтобы показать некоторые из содержащихся в ней идей.

Рассмотрим уравнение

$$y'' = f(x)y + g(x), \quad y(0) = A, \quad y(1) = B. \quad (16.7-1)$$

Прежде всего аппроксимируем y'' второй разностью

$$h^2 y_n'' = \Delta^2 y_{n-1}, \quad (16.7-2)$$

считая, что интервал $(0 \leq x \leq 1)$ разделен на N интервалов величиной $h = \frac{1}{N}$. Таким образом,

$$y_{n+1} - 2y_n + y_{n-1} = h^2 [f(x_n)y_n + g(x_n)]$$

или

$$y_n = \frac{y_{n-1} + y_{n+1} - h^2 g(x_n)}{2 + h^2 f(x_n)} \quad (n = 1, 2, \dots, N-1). \quad (16.7-3)$$

Получили $N-1$ линейное уравнение с $N-1$ неизвестным, которые могут быть решены многими различными способами.

В качестве примера рассмотрим один частный случай уравнения (16.7-1). Положим $A=B=0$, $f(x)=1$, $g(x)=x$ и $N=4$, т. е. уравнение

$$y'' = y + x, \quad y(0) = 0, \quad y(1) = 0,$$

Требуемое решение известно:

$$y = \frac{\operatorname{sh} x}{\operatorname{sh} 1} - x.$$

Разностные уравнения имеют вид

$$\begin{aligned} y_2 - 2y_1 &= \frac{1}{16} \left(y_1 + \frac{1}{4} \right), \\ y_3 - 2y_2 + y_1 &= \frac{1}{16} \left(y_2 + \frac{2}{4} \right), \\ -2y_2 + y_3 &= \frac{1}{16} \left(y_3 + \frac{3}{4} \right), \end{aligned}$$

или

$$\begin{aligned} 16y_2 - 33y_1 &= \frac{1}{4}, \\ 16y_3 - 33y_2 + 16y_1 &= \frac{2}{4}, \\ -33y_2 + 16y_3 &= \frac{3}{4}. \end{aligned}$$

Эти уравнения можно решить, например, так. Умножим первое из уравнений на 16, второе на 33 и третье на 16 и сложим. Получим

$$(16^2 - 33^2 + 16^2)y_2 = \frac{1}{4}(16 + 2 \cdot 33 + 3 \cdot 16).$$

Зная y_2 , легко найти y_1 и y_3 из первого и третьего уравнений. Но что делать, если полученное решение недостаточно точно? Можно положить $N=8$ и, используя вычисленное решение для $N=4$, прикинуть решения для $N=8$. Затем, подставляя это решение в правые части восьми уравнений, соответствующих (16.7-3), можно вычислить улучшенные значения и повторять процесс до тех пор, пока не прекратятся изменения. Число точек можно увеличивать как угодно.

Наоборот, вместо (16.7-2), наверное, лучше было бы взять более точную формулу. Чтобы найти такую формулу, используем матрицу S_7 из § 10.3. Моменты $y''(0)$ суть $(0, 0, 2, 0, 0, 0, 0)$ и могли бы быть записаны как вектор-столбец. Это дает вектор-столбец, соответствующий удвоенному третьему столбцу S_7 :

$$\frac{1}{360} \begin{pmatrix} 4 \\ -54 \\ 540 \\ -980 \\ 540 \\ -54 \\ 4 \end{pmatrix} = \frac{1}{180} \begin{pmatrix} 2 \\ -27 \\ 270 \\ -490 \\ 270 \\ -27 \\ 2 \end{pmatrix},$$

который приводит к формуле

$$h^2 y_n'' = \frac{1}{180} (2y_{n+3} - 27y_{n+2} + 270y_{n+1} - 490y_n + 270y_{n-1} - \\ - 27y_{n-2} + 2y_{n-3}) = \Delta^3 y_{n-1} - \frac{1}{12} \Delta^4 y_{n-2} + \frac{1}{90} \Delta^6 y_{n-3}.$$

Беря в ней только первые два члена, так как величина $\frac{1}{90} \Delta^6 y_{n-3}$, вероятно, должна быть маленькой, найдем значение $-\frac{1}{12} \Delta^4 y_{n-1}$. Чтобы вычислить его, нужны несколько значений вне интервала ($0 \leq x \leq 1$); в частности, надо знать $y(-\frac{1}{4})$ и $y(\frac{5}{4})$. Эти величины могут быть найдены из очевидных соотношений:

$$y_{-1} - 2y_0 + y_1 = h^2(y_0 + 0), \quad y_3 - 2y_4 + y_5 = h^2(y_4 + 1).$$

Заметим, что в действительности нам нужны лишь $\Delta^3 y_{-1}$ и $\Delta^3 y_3$, которые могут быть найдены из значений y_0 и y_4 : $\Delta^3 y_{-1} = 0$, $\Delta^3 y_4 = h^3$. Используя таблицу значений

	Δ	Δ^2	Δ^3	Δ^4
y_{-1}	—			
y_0	—	—		
y_1	—	—	—	—
y_2	—	—	—	—
y_3	—	—	—	—
y_4	—	—		
y_5				

можно вычислить значения $\Delta^4 y_i$, а именно $\Delta^4 y_{-1}$, $\Delta^4 y_0$, $\Delta^4 y_1$, требующиеся, чтобы улучшить результат.

Эти значения $\Delta^4 y_i$, которые не учитывались в первом вычислении подставим теперь в корректирующие члены в правой части

$$\Delta^3 y_{n-1} = h^3(y_n + x_n) + \frac{1}{12} \Delta^4 y_{n-3}.$$

Решим задачу, опять не меняя членов $\Delta^4 y_i$. Если значения $\Delta^4 y_i$ нового решения существенно отличаются от старых, то повторим процесс опять. Фокс [11], который использовал описанный прием с большой эффективностью, назвал его методом «разности и коррекции»

Как уже было замечено, мы не пытаемся излагать тщательно разработанную теорию для краевых задач, а лишь стремимся дать простейшие приемы, показывающие, как с такими задачами обращаться.

Упражнение 16.7-1. Выполнить вычисления примера в § 16.7, используя метод разности и коррекции, и сравнить результат с правильным ответом.

ГЛАВА 17

МЕТОД НАИМЕНЬШИХ КВАДРАТОВ. ТЕОРИЯ

§ 17.1. Введение

В § 7.1 ставились четыре основных вопроса:

1. Какие узловые точки мы будем использовать?
2. Какой класс аппроксимирующих функций мы будем использовать?
3. Какой критерий согласия мы примем?
4. Какую точность мы хотим иметь?

До сих пор было рассмотрено много различных методов для выбора узлов, но мы всегда пользовались многочленами и критерием точного прохождения многочлена через узловые точки. Проведем теперь исследование других критериев для выбора конкретного многочлена из общего класса всех многочленов степени n .

Окончательный выбор критерия для конкретной задачи зависит от предыстории данных уравнений и поэтому не может быть дан в учебнике по методам вычислений. Попробуем, однако, обсудить некоторые критерии, рассмотрев их достаточно подробно.

Когда хорош критерий точного соответствия? Известно, что значения функции в выбранных точках искажены шумом округления. Пока уровень шума слабый, а во многих расчетах на современных вычислительных машинах с 8—12 десятичными разрядами это так, точное соответствие является разумным методом.

Если же уровень шума высок, что может случиться в некоторых вычислениях — и почти всегда случается, когда данные получены из физических измерений, — то разумность попытки искать аппроксимирующую функцию по критерию точного совпадения в «шумных» узлах необходимо исследовать внимательно. По-видимому, наиболее широко использующийся метод в «шумных» ситуациях — это аппроксимация по наименьшим квадратам.

§ 17.2. Метод наименьших квадратов

Предположим, что требуется измерить некоторую величину и делается n измерений, результаты которых равны

$$x_i = x + \varepsilon_i \quad (i = 1, 2, \dots, n),$$

где ε_i — это ошибки (или шум) измерений, а x — это «истинное значение».

Метод наименьших квадратов утверждает, что наилучшее приближенное значение \bar{x} есть такое число, для которого минимальна *) сумма квадратов отклонений от \bar{x}_i :

$$f(\bar{x}) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (x_i - \bar{x})^2$$

В конечном счете, полезность этого метода определяется тем, насколько хорошо эта модель соответствует опыту и как легко можно использовать ее на практике. Стоит заметить, что этот метод эквивалентен предположению, что наилучшим приближением является среднее арифметическое

$$x_a = \frac{1}{n} \sum_{i=1}^n x_i.$$

Чтобы доказать эквивалентность этих определений, покажем сначала, что метод наименьших квадратов приводит к среднему арифметическому. Заметим, что

$$f(\bar{x}) = \sum_i (x_i - \bar{x})^2$$

можно рассматривать как функцию от \bar{x} и минимизировать обычным приемом:

$$\begin{aligned} \frac{df}{d\bar{x}} = -2 \sum_i (x_i - \bar{x}) = 0, \quad \sum_i x_i - \sum_i \bar{x} = 0, \quad \sum_i x_i - n\bar{x} = 0, \\ \bar{x} = \frac{1}{n} \sum_i x_i = x_a. \end{aligned}$$

Таким образом, $\bar{x} = x_a$ минимизирует $\sum (x_i - \bar{x})^2$, так как

$$\frac{d^2 f}{d\bar{x}^2} = 2n > 0.$$

Это же можно показать и иначе. Пусть

$$f(x_a) = \sum (x_i - x_a)^2.$$

Тогда

$$\begin{aligned} f(x_a) = \sum x_i^2 - 2x_a \sum x_i + \sum x_a^2 = \sum x_i^2 - 2x_a n x_a + n x_a^2 = \\ = \sum x_i^2 - n x_a^2. \end{aligned}$$

*) Теорема Гаусса — Маркова утверждает, что если среднее по всем ε_i и корреляция по всем $\varepsilon_i \varepsilon_j$ равны нулю, то эта оценка имеет наименьшую дисперсию из всех линейных оценок.

Если взять любое значение x_b , отличное от x_a то

$$f(x_b) = \sum (x_i - x_b)^2 = \sum x_i^2 - 2nx_ax_b + nx_b^2.$$

В таком случае имеем

$$f(x_b) - f(x_a) = n(x_a^2 - 2x_ax_b + x_b^2) = n(x_a - x_b)^2 \geq 0.$$

Таким образом, сумма квадратов \bar{s}_i минимальна только при $\bar{x} = x_a$. Следовательно, доказано, что «метод наименьших квадратов» и «выбор среднего» эквивалентны.

Упражнение 17.2-1. Найти приближенное значение в смысле наименьших квадратов для чисел 2, 3, 2, 1, 2, 3 обоими способами.

О т в е т: $\bar{x} = 13/6$.

§ 17.3. Другие критерии

Выбор в качестве лучшего значения среднего из n измерений соответствует методу наименьших квадратов. Но возможны и другие методы, и часто они более уместны.

Предположим, что вместо минимизации суммы квадратов требуется минимизация суммы модулей отклонений

$$\sum_i |x_i - \bar{x}| = f(\bar{x}).$$

Это требование приводит к выбору медианного (срединного) значения x_m из x_i (если количество x_i четно, то приходим к выбору любого из двух срединных). Доказательство этого факта простое. Предположим, что имеется нечетное число $2k+1$ значений x_i . Выберем в качестве x_m среднее из них по величине. Тогда любой сдвиг от x_m по x , скажем вверх, будет увеличивать k членов $|x_i - x|$, для которых x_i ниже x_m , и уменьшать k членов $|x_i - x|$, для которых x_i выше x_m каждый на одинаковую величину; но член $|x_m - x|$ будет также увеличиваться, увеличивая, таким образом, всю сумму отклонений.

Другой метод состоит в минимизации максимального отклонения вместо минимизации суммы их квадратов. Это приводит к значению

$$\frac{x_{\max} + x_{\min}}{2} = x_{\text{ср}},$$

которое является серединой интервала x_i .

§ 17.4. Ошибки с нормальным распределением

По-видимому, широко распространено мнение, что метод наименьших квадратов подразумевает распределение ошибок по нормальному закону. Согласно ему, вероятность того, что ошибка ε_i находится

в интервале $x, x + \Delta x$, задается формулой

$$\frac{k}{\sqrt{\pi}} e^{-k^2 x^2} \Delta x. \quad (17.4-1)$$

Это мнение ошибочно, и здесь стоит обсудить эту тему, чтобы прояснить дело. Кроме того, иногда верят в то, что нормальный закон есть закон природы. Распространена формулировка: «Математики полагают, что нормальный закон должен быть физическим законом, в то время как физики думают, что он должен быть математическим законом».

Одна из характерных ситуаций, приводящих к нормальному закону (Гершель), выглядит так: рассмотрим метание дротика с некоторой высоты с целью попасть в некоторую точку O на горизонтальном полу (рис. 17.4-1). Предположим теперь, что ошибки не зависят от выбора системы координат, а зависят только от расстояния до точки O и что большие ошибки менее вероятны, чем меньшие. Эти предположения кажутся вполне разумными, так как система координат может быть вполне произвольной.

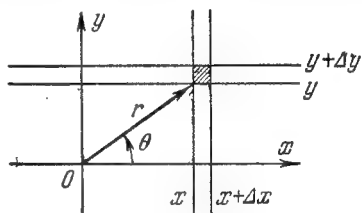


Рис. 17.4-1.

Пусть вероятность попадания в полосу $x, x + \Delta x$ равна (приближенно) $f(x) \Delta x$ и соответственно в полосу $y, y + \Delta y$ равна $f(y) \Delta y$. Исходя из предположения о независимости ошибок по каждой координате, получаем, что вероятность попадания в прямоугольник пересечения двух полос равна

$$f(x) f(y) \Delta x \Delta y. \quad (17.4-2)$$

Введем теперь полярные координаты. Вероятность попадания в элемент площади по предположению не зависит от направления и равна

$$g(r) \Delta x \Delta y. \quad (17.4-3)$$

Так как (17.4-2) и (17.4-3) выражают одну и ту же вероятность, то

$$g(r) = f(x) f(y). \quad (17.4-4)$$

Левая часть не зависит от θ , и, дифференцируя по θ , имеем

$$\frac{\partial g(r)}{\partial \theta} = 0 = f(x) \frac{\partial f(y)}{\partial \theta} + f(y) \frac{\partial f(x)}{\partial \theta}. \quad (17.4-5)$$

Используя соотношения $x = r \cos \theta$, $y = r \sin \theta$, находим

$$\left. \begin{aligned} \frac{\partial f(x)}{\partial \theta} &= \frac{df(x)}{dx} \frac{\partial x}{\partial \theta} = f'(x) \cdot (-y), \\ \frac{\partial f(y)}{\partial \theta} &= \frac{df(y)}{dy} \frac{\partial y}{\partial \theta} = f'(y) \cdot x. \end{aligned} \right\} \quad (17.4-6)$$

Подставляя (17.4-6) в (17.4-5), получаем

$$f(x)f'(y)x + f(y)f'(x)(-y) = 0$$

или

$$\frac{f'(x)}{xf(x)} = \frac{f'(y)}{yf(y)}.$$

По предположению x и y независимы, а значит, обе части этого равенства равны константе, скажем, K , т. е.

$$\frac{f'(x)}{xf(x)} = K = \frac{f'(y)}{yf(y)}$$

или, что то же,

$$\frac{df(x)}{f(x)} = K \cdot x \cdot dx \quad \text{и} \quad \frac{df(y)}{f(y)} = K \cdot y \cdot dy.$$

Интегрируя первое из этих равенств, получаем

$$\ln f(x) = \frac{Kx^2}{2} + C \quad \text{или} \quad f(x) = Ae^{\frac{1}{2}Kx^2}.$$

Мы предположили, что бóльшие ошибки менее вероятны, чем меньшие; поэтому K должно быть отрицательным, скажем, $K = -2k^2$.

Окончательно имеем

$$f(x) = Ae^{-k^2x^2}, \quad f(y) = Ae^{-k^2y^2}$$

и из (17.4-4)

$$g(r) = A^2e^{-k^2(x^2+y^2)}.$$

Но, так как куда-нибудь дротик должен попасть, то

$$\int_0^{2\pi} \int_0^\infty g(r) r dr d\theta = 1, \quad A^2 2\pi \int_0^\infty r e^{-k^2r^2} dr = 1,$$

$$A^2 \pi \frac{e^{-k^2r^2}}{-k^2} \Big|_0^\infty = \frac{A^2 \pi}{k^2} = 1, \quad A = \frac{k}{\sqrt{\pi}},$$

и окончательно мы получаем (формула (17.4-1))

$$f(x) = \frac{k}{\sqrt{\pi}} e^{-k^2x^2},$$

т. е. нормальное (гауссово) распределение.

Заметим, что это не единственный способ получить нормальный закон. Другой подход дает центральная предельная теорема, которая утверждает, грубо говоря, что сумма большого числа маленьких ошибок распределена нормально.

На практике нормальный закон есть обычная модель во многих приложениях. Отклонение от него происходит обычно от наличия

большого, чем это предсказывается моделью, количества попадания в «хвост» распределения, где $|x|$ велик. Причиной этого часто является небольшой «размазанный» эффект. В таких случаях обычна смесь двух нормальных кривых с различными значениями k .

Теория качественного контроля, в частности, основана на возмущениях, наблюдаемых в «хвостах».

Упражнение 17.4-1. Показать, что $\sigma^2 = \frac{V_2}{k^2}$, где σ есть дисперсия распределения $f(x) = \frac{k}{\sqrt{\pi}} e^{-k^2 x^2}$.

§ 17.5. Проведение подходящего многочлена

Один из наиболее общих случаев применения метода наименьших квадратов состоит в том, что имеется N наблюдений (x_i, y_i) ($i = 1, 2, \dots, N$) и требуется приблизить эти данные многочленом степени $M < N$,

$$y(x) = a_0 + a_1 x + \dots + a_M x^M. \quad (17.5-1)$$

Вычисленная кривая $y(x)$ в некотором смысле дает сглаженное множество значений $y(x_i)$, которые, вообще говоря, отличны от наблюдаемых y_i . Метод наименьших квадратов утверждает, что следует выбирать такой многочлен, который минимизирует функцию

$$\sum_{i=1}^N [y_i - y(x_i)]^2 = f(a_0, a_1, \dots, a_M). \quad (17.5-2)$$

До сих пор рассматривался метод наименьших квадратов только для одного измерения; очевидно, что принципиально ничего не меняется при переходе к $(M+1)$ измерению. Можно рассматривать a_0, a_1, \dots, a_M как координаты одной точки в $(M+1)$ -мерном евклидовом пространстве.

Для нахождения минимума поступаем как в анализе: дифференцируем (17.5-2) по каждой из неизвестных a_k

$$\frac{\partial f}{\partial a_k} = -2 \sum_i [y_i - y(x_i)] x_i^k = 0 \quad (k = 0, 1, 2, \dots, M)$$

или

$$\begin{aligned} \sum_i y_i x_i^k &= a_0 \sum_i x_i^k + a_1 \sum_i x_i^{k+1} + \dots + a_M \sum_i x_i^{k+M} = \\ &= \sum_{i=0}^M a_i \sum_{i=1}^N x_i^{k+i}. \end{aligned} \quad (17.5-3)$$

Для упрощения введем обозначения

$$\sum_{i=1}^N x_i^k = S_k, \quad \sum_{i=1}^N y_i x_i^k = T_k$$

Уравнения (17.5-3) принимают вид

$$\sum_{j=0}^M a_j S_{k+j} = T_k \quad (k=0, 1, \dots, M) \quad (17.5-4)$$

и называются «нормальными уравнениями».

Они образуют систему $M+1$ линейных уравнений, определитель которых есть $\Delta = |S_{k+j}|$.

Покажем, что $\Delta \neq 0$. Если бы $\Delta = 0$, то однородная система, соответствующая (17.5-4),

$$\sum_{j=0}^M a_j S_{k+j} = 0$$

имела бы ненулевое решение. Умножая k -е из этих уравнений на a_k и суммируя по всем k , получаем

$$\begin{aligned} 0 &= \sum_{k=0}^M a_k \sum_{j=0}^M a_j \sum_{i=1}^N (x_i^j x_i^k) = \sum_{i=1}^N \left(\sum_{k=0}^M a_k x_i^k \right) \left(\sum_{j=0}^M a_j x_i^j \right) = \\ &= \sum_{i=1}^N \left(\sum_{k=0}^M a_k x_i^k \right)^2 = \sum_{i=1}^N y^2(x_i) = 0, \end{aligned}$$

откуда все $y(x_i) = 0$, $i = 1, 2, \dots, N$. Так как $N > M$, это невозможно вследствие фундаментальной теоремы алгебры: многочлен степени M имеет не больше M корней, если не все $a_k = 0$. Таким образом, не существует ненулевого решения и $\Delta \neq 0$.

В принципе наша задача решена. На практике же решать систему (17.5-4) не так легко, потому что определитель Δ часто бывает весьма близок к нулю. Чтобы увидеть, как это может случиться, предположим, что x_i более или менее равномерно распределены в интервале $0 \leq x_i \leq 1$. Тогда

$$S_k = \sum_{i=1}^N x_i^k \approx N \int_0^1 x^k dx = \frac{1}{k+1} N.$$

Определитель (с точностью до множителя N^{M+1})

$$\left| \frac{1}{i+j+1} \right| \quad (i, j = 0, 1, \dots, M)$$

известен как *определитель Гильберта*.

Определитель Гильберта порядка n имеет величину

$$H_n = \frac{[1! \ 2! \ 3! \ \dots \ (n-1)!]^2}{n! \ (n+1)! \ \dots \ (2n-1)!},$$

которая быстро стремится к нулю. В таблице 17.5-1 приведено несколько значений H_n .

Таблица 17.5-1

Значения определителей Гильберта

n	H_n	n	H_n	n	H_n
1	1	4	$1,7 \cdot 10^{-7}$	7	$4,8 \cdot 10^{-25}$
2	$8,3 \cdot 10^{-2}$	5	$3,7 \cdot 10^{-12}$	8	$2,7 \cdot 10^{-33}$
3	$4,6 \cdot 10^{-4}$	6	$5,4 \cdot 10^{-16}$	9	$9,7 \cdot 10^{-43}$

Чтобы обойти трудности решения системы с очень малым определителем, стоит вычислять не S_k , а некоторый их эквивалент. Это приводит к ортогональным функциям и специальным ортогональным полиномам, которые хорошо изучены и по которым имеется обширная литература.

С другой стороны, заметим, что уравнения (17.5-4) могут быть записаны в других обозначениях. Из (17.5-1) имеем

$$a_0 + a_1 x_i + a_2 x_i^2 + \dots + a_M x_i^M = y(x_i) \quad (i = 1, 2, \dots, N)$$

или, в матричном обозначении, $X_a = y$. Пусть X^T — транспозиция матрицы X . Тогда (17.5-4) принимает вид

$$X^T X_a = X^T y.$$

Этот путь на практике не пригоден.

Упражнение 17.5-1. Дано

x	0	1	2	3	4
y	1	2	1	0	4

Провести прямую методом наименьших квадратов.

$$\text{Ответ: } y = \frac{2}{5}(x + 2).$$

§ 17.6. Ортогональные функции

В случае двух измерений две прямые называются «ортогональными» или перпендикулярными, если

$$\operatorname{tg} \theta_1 = -\frac{1}{\operatorname{tg} \theta_2}.$$

Это условие можно записать в виде

$$\sin \theta_1 \sin \theta_2 + \cos \theta_1 \cos \theta_2 = 0.$$

Используя дополнительные углы φ_1 и φ_2 в первом слагаемом, получаем

$$\cos \varphi_1 \cos \varphi_2 + \cos \theta_1 \cos \theta_2 = 0.$$

В случае трех измерений с углами α , β , γ в качестве условия перпендикулярности или ортогональности имеем

$$\cos \alpha_1 \cos \alpha_2 + \cos \beta_1 \cos \beta_2 + \cos \gamma_1 \cos \gamma_2 = 0.$$

Условие ортогональности легко обобщить на случай n измерений с углами $\alpha^{(1)}$, $\alpha^{(2)}$, ..., $\alpha^{(n)}$

$$\cos \alpha_1^{(1)} \cos \alpha_2^{(1)} + \dots + \cos \alpha_1^{(n)} \cos \alpha_2^{(n)} = 0.$$

На практике оказалось, что направляющие косинусы

$$\cos \alpha_i^{(k)} = \lambda_i(k)$$

более удобны, чем сами углы, так что обычно условие ортогональности употребляется в форме

$$\sum_{k=1}^n \lambda_1(k) \cdot \lambda_2(k) = 0.$$

Если теперь в некотором смысле устремить $k \rightarrow \infty$, то можно заметить, что разумно назвать условием ортогональности равенство

$$\int \lambda_1(k) \lambda_2(k) dk = 0.$$

Исходя из этого, говорят, что две функции $f_1(x)$ и $f_2(x)$ ортогональны на интервале $a \leq x \leq b$, если

$$\int_a^b f_1(x) f_2(x) dx = 0.$$

Функции f_1, f_2, \dots, f_m взаимно ортогональны, если

$$\left. \begin{aligned} \int_a^b f_i(x) f_j(x) dx &= 0 & (i \neq j), \\ \int_a^b f_i^2(x) dx &= \lambda_i > 0 & (i = j) \end{aligned} \right\} \quad (17.6-1)$$

(к сожалению, в этом месте обычно употребляют обозначения λ_i ; их не следует путать с направляющими косинусами).

Мы предполагаем, что $f_i(x)$ действительны, непрерывны и не равны тождественно нулю.

Пусть

$$g_i = \frac{f_i}{\sqrt{\lambda_i}}.$$

Тогда

$$\int_a^b g_i(x) g_j(x) dx = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases} \quad (17.6-2)$$

Функции $g_i(x)$ называются *ортонормированными*, а процесс перехода от f_i к g_i — *нормированием*.

Классической системой ортогональных функций в интервале $0 \leq x < 2\pi$ (или $-\pi \leq x \leq \pi$) является система

$$\begin{aligned} 1, \cos x, \cos 2x, \dots, \cos Mx, \\ \sin x, \sin 2x, \dots, \sin Mx. \end{aligned}$$

Если значения заданы в $2N$ равноотстоящих точках, ряды косинусов и синусов заканчиваются соответственно $\frac{1}{2} \cos Nx$ и $\sin (N-1)x$ (см. гл. 6).

Чтобы увидеть, как метод ортогональных функций обходит трудности, порожденные плохой матрицей неизвестных коэффициентов, напомним, как производится разложение. Требуется представить y_i в виде

$$\begin{aligned} y_i = \frac{a_0}{2} + a_1 \cos x_i + \dots + a_{N-1} \cos (N-1)x_i + \frac{a_N}{2} \cos Nx_i + \\ + b_1 \sin x_i + \dots + b_{N-1} \sin (N-1)x_i. \end{aligned}$$

Умножая на $\cos mx_i$ или $\sin mx_i$ и суммируя по всем $x_i = \frac{\pi i}{N}$, получаем (используя (6.2-3))

$$a_m = \frac{1}{N} \sum_{i=0}^{2N-1} y_i \cos mx_i, \quad b_m = \frac{1}{N} \sum_{i=0}^{2N-1} y_i \sin mx_i.$$

Эта система уравнений для неизвестных a_m и b_m решается тривиально — она уже решена.

Все происходит так же, если есть другое множество ортогональных функций $\{f_i(x)\}$. Пусть $f(x) = a_0 f_0(x) + \dots + a_N f_N(x)$. Умножаем на $f_j(x)$ и интегрируем:

$$\int_a^b f(x) f_j(x) dx = \lambda_j a_j \quad (17.6-3)$$

Решение также тривиально. Коэффициенты a_j , найденные таким способом, называются *коэффициентами Фурье*.

Если функция известна только в отдельных точках x_m , то интегралы заменяются суммами, а условия ортогональности принимают вид

$$\sum_m f_i(x_m) f_j(x_m) = \begin{cases} 0 & \text{при } j \neq i, \\ \lambda_i & \text{при } j = i. \end{cases}$$

Заметим, что $\lambda_j \neq 0$. Если данные имеют различную степень важности, в интеграл может быть введена весовая функция $\rho(x) \geq 0$

$$\int_a^b \rho(x) f_i(x) f_j(x) dx, \quad (17.6-4)$$

а коэффициенты получаются из уравнений

$$\int_a^b \rho(x) f(x) f_j(x) dx = \lambda_j a_j \quad (17.6-5)$$

Если задана смешанная информация, в непрерывной и дискретной формах, то можно использовать интеграл Стильтьеса.

Развить эту теорию легко для непрерывной формы, хотя большей частью она будет нужна для дискретного случая.

Упражнение 17.6-1. Показать, что $P_0 = 1$; $P_1 = x$; $P_2 = \frac{1}{2}(3x^2 - 1)$; $P_3 = \frac{1}{2}(5x^3 - 3x)$ ортогональны для $-1 \leq x \leq 1$.

§ 17.7. Общие свойства ортогональных функций

Прежде чем пользоваться ортогональными функциями, разберем некоторые общие теоремы, чтобы дать читателю представление о поведении ортогональных функций.

Идея линейной независимости является одной из основных идей в математике. Функции $f_i(x)$ называются *линейно независимыми* на (a, b) , если из равенства $a_0 f_0(x) + a_1 f_1(x) + \dots + a_m f_m(x) \equiv 0$ во всем

интервале следует, что все $a_i = 0$ *); в противном случае они линейно зависимы.

Простым и важным примером множества линейно независимых функций в произвольном интервале является множество функций

$$1, x, x^2, x^3, \dots, x^m,$$

так как по основной теореме алгебры из тождества

$$a_0 + a_1x + \dots + a_mx^m \equiv 0$$

следует, что все $a_i = 0$.

Функции непрерывные и ортогональные в интервале линейно независимы. Доказательство тривиально: допустим, что $a_0f_0 + a_1f_1 + \dots + a_mf_m \equiv 0$, и вычислим коэффициенты Фурье (17.6-5)

$$a_j = \frac{1}{\lambda_j} \int_a^b 0 \cdot \rho(x) f_j(x) dx = 0.$$

Наоборот, из системы линейно независимых функций с помощью процесса Шмидта можно получить систему ортогональных функций. Процесс состоит в следующем. Пусть дано множество линейно независимых функций $f_i(x)$. Вычислим

$$\int_a^b \rho(x) f_0^2(x) dx = \lambda_0 > 0 \quad [\rho(x) \geq 0].$$

Тогда равенство

$$g_0(x) = \frac{f_0(x)}{\sqrt{\lambda_0}}$$

определяет первую ортонормированную функцию $g_0(x)$. Применяя метод математической индукции, предположим, что уже построены первые j ортонормированных функций $g_i(x) (i=0, 1, 2, \dots, j-1)$.

Положим

$$F_j(x) = a_0g_0 + a_1g_1 + \dots + a_{j-1}g_{j-1} + f_j(x). \quad (17.7-1)$$

Функция $F_j(x)$ отлична от нуля, так как $f_i(x)$ линейно независимы, а каждая $g_i(x)$ есть линейная комбинация $f_k(x)$ для $k \leq i$. Мы должны иметь

$$\int_a^b \rho(x) F_j(x) g_i(x) dx = 0, \quad 0 \leq i \leq j-1.$$

*) Функцию $f(x) \equiv 0$ здесь и дальше исключаем из рассмотрения.

Но по определению $F_j(x)$ отсюда следует

$$\lambda_i a_i + \int_a^b \rho(x) g_i(x) f_j(x) dx = 0,$$

откуда находится a_i и тем самым $F_j(x)$. Чтобы нормировать $F_j(x)$ нужно вычислить

$$\int_a^b \rho(x) F_j^2(x) dx = \lambda_j \quad [\rho(x) \geq 0]$$

и затем положить $g_j(x) = \frac{F_j(x)}{\sqrt{\lambda_j}}$.

Таким образом, шаг индукции выполнен. Если есть только конечное число N узлов x_m , то существует по крайней мере N линейно независимых функций $f_j(x_m)$.

То, что их N , следует из существования множества

$$g_j(x_m) = \begin{cases} 0 & \text{при } m \neq j, \\ 1 & \text{при } m = j, \end{cases} \quad (j = 1, \dots, N),$$

так как никакое множество этих функций не может быть линейно зависимым.

К сожалению, семейство функций, которое получается в этом процессе, определяется неоднозначно. Оно зависит от выбора $\sqrt{\lambda_i} > 0$, а также от интервала, порядка, в котором мы перебираем функции, и весовой функции $\rho(x)$.

Упражнение 17.7-1. Даны $P_0 = 1$; $P_1 = x$; $\rho(x) = 1$.

Построить ортогональные функции $P_2(x)$ и $P_3(x)$ на множестве точек $-2, -1, 0, 1, 2$, где P_2 и P_3 — многочлены степени 2 и 3 соответственно.

§ 17.8. Неравенство Бесселя и полнота

Коэффициенты Фурье функции $F(x)$ относительно ортонормированного семейства

$$a_j = \int_a^b \rho(x) F(x) g_j(x) dx$$

удовлетворяют неравенству Бесселя

$$\int_a^b \rho(x) F^2(x) dx \geq \sum_{j=0}^M a_j^2 \quad (\rho \geq 0). \quad (17.8-1)$$

Доказывается это непосредственно. Напишем

$$\int_a^b \rho(x) \left[F(x) - \sum_{i=0}^M a_i g_i(x) \right]^2 dx \geq 0$$

и раскроем скобки

$$\begin{aligned} 0 \leq \int_a^b \rho(x) F^2(x) dx - 2 \int_a^b \rho(x) F(x) \sum_{i=0}^M a_i g_i(x) dx + \\ + \int_a^b \sum_{i=0}^M \sum_{j=0}^M a_i a_j \rho(x) g_i g_j dx. \end{aligned}$$

Используя определение a_i и ортогональность g_i , находим

$$\int_a^b \rho(x) F^2(x) dx \geq 2 \sum_{i=0}^M a_i^2 - \sum_{i=0}^M a_i^2 = \sum_{i=0}^M a_i^2$$

для всех M .

В непрерывном случае, если в (17.8-1) выполняется равенство для любой функции $F(x)$, непрерывной в (a, b) , то бесконечное множество функций $g_i(x)$ называется *полным* *), а равенство

$$\int_a^b \rho(x) F^2(x) dx = \sum_{i=0}^{\infty} a_i^2 \quad (17.8-2)$$

называется *равенством Парсеваля*.

В дискретном случае у нас не возникает трудностей при перестановке суммирования, и любое семейство N функций, линейно независимых на множестве N точек, является полным на этом множестве точек (ср. § 6.3).

§ 17.9. Метод наименьших квадратов и коэффициенты Фурье

Коэффициенты Фурье a_j дают наилучшее в смысле наименьших квадратов приближение, когда $F(x)$ разлагается по ортогональному множеству функции $g_j(x)$. Чтобы доказать это,

*) Это свойство чаще называют замкнутостью системы функций, а под полнотой множества понимается отсутствие функции, ортогональной всем функциям множества. Впрочем, для функций с интегрируемым квадратом эти свойства эквивалентны. (Прим. ред.)

минимизируем выражение

$$\begin{aligned}
 m &= \int_a^b \rho(x) \left[F(x) - \sum_{j=0}^M c_j g_j(x) \right]^2 dx = \int \rho(x) F^2(x) dx - \\
 &\quad - 2 \sum_{j=0}^M c_j \int \rho(x) F(x) g_j(x) dx + \sum_{i=0}^M \sum_{j=0}^M c_i c_j \int \rho(x) g_i g_j dx = \\
 &= \int \rho(x) F^2(x) dx - 2 \sum_{i=0}^M c_i a_i + \sum_{i=0}^M c_i^2 = \\
 &= \int \rho(x) F^2(x) dx - \sum_{i=0}^M a_i^2 + \sum_{i=0}^M (a_i - c_i)^2.
 \end{aligned}$$

Но последнее выражение минимально, если $c_i = a_i$, что и требовалось.

Мы получили замечательное и очень полезное свойство коэффициентов Фурье: каждый из коэффициентов a_i , дающий лучшее приближение в смысле наименьших квадратов по системе ортогональных функций, определяется независимо от остальных, и если требуется изменить число используемых функций $g_i(x)$, то не нужно искать заново уже найденные коэффициенты. Исследуем теперь обратную задачу.

Если коэффициенты c_i разложения функции $F(x)$ по множеству функции $\mu_i(x)$ в смысле наименьших квадратов не изменяются при изменении числа используемых $\mu_i(x)$, то $\mu_i(x)$ должны быть ортогональны. Положим

$$g(c_0, c_1, \dots, c_n) = \int_a^b \rho(x) \left[F(x) - \sum_{i=0}^M c_i \mu_i(x) \right]^2 dx.$$

Так как g должно быть минимизировано, то

$$\frac{\partial g}{\partial c_j} = 0 = -2 \int_a^b \rho(x) \left[F(x) - \sum_{i=0}^M c_i \mu_i(x) \right] \mu_j(x) dx$$

или

$$\int_a^b \rho(x) F(x) \mu_j dx = \sum_{i=0}^M c_i \int_a^b \rho(x) \mu_i \mu_j dx. \quad (17.9-1)$$

Если это свойство верно для всех M , оно должно быть верно и для $(M+1)$

$$\int_a^b \rho(x) F(x) \mu_j(x) dx = \sum_{i=0}^{M+1} c_i \int_a^b \rho(x) \mu_i \mu_j dx. \quad (17.9-2)$$

Из (17.9-1) и (17.9-2) имеем

$$c_{M+1} \int_a^b \rho(x) \mu_{M+1} \mu_j dx = 0$$

для любого j , т. е. μ_j ортогонально μ_{M+1} (а M было произвольно).

Таким образом, ортогональные функции, нахождение коэффициентов Фурье и идея приближения в смысле наименьших квадратов тесно переплетаются.

§ 17.10. Ортогональные многочлены

Важным подклассом ортогональных функций является подкласс ортогональных многочленов, где k -й многочлен имеет степень k ($k = 0, 1, \dots$).

Легко показать, что k -й ортогональный многочлен y_k имеет k действительных различных корней в интервале интегрирования. Действительно, предположим, что число различных действительных корней $r < k$. образуем произведение

$$\pi(x) = (x - x_1)(x - x_2) \dots (x - x_r) \quad (r < k).$$

Тогда

$$\int_a^b \rho(x) \pi(x) y_k(x) dx = 0 \quad (\rho \geq 0),$$

так как $\pi(x)$ может быть разложена по $y_0 y_1 \dots y_r$, а $y_k(x)$ ортогонально им всем. Но это невозможно, поскольку подынтегральная функция не меняет знак в интервале. Следовательно, на (a, b) существует k действительных различных корней.

Ортогональные многочлены $y_k(x)$ удовлетворяют трехчленному рекуррентному соотношению вида

$$a_k y_{k+1}(x) + (b_k - x) y_k(x) + c_k y_{k-1}(x) = 0 \quad (k \geq 1). \quad (17.10-1)$$

Чтобы показать это, положим

$$y_i(x) = \alpha_i x^i + \dots \quad (\alpha_i > 0).$$

(Выбор α_i — некоторого положительного числа — есть дело только соглашения.)

Тогда разность

$$a_k y_{k+1} - x y_k$$

есть многочлен степени k при условии, что $a_k = \frac{\alpha_k}{\alpha_{k+1}}$. Следовательно,

$$a_k y_{k+1} - x y_k = \gamma_k y_k + \gamma_{k-1} y_{k-1} + \dots + \gamma_0 y_0 \quad (17.10-2)$$

Умножим это равенство на $\rho(x) y_m(x)$ и проинтегрируем. Получаем

$$\int_a^b \rho(x) (a_k y_{k+1} - x y_k) y_m(x) dx = \gamma_m \lambda_m. \quad (17.10-3)$$

Для $m=0, 1, \dots, k$ функция y_m ортогональна y_{k+1} , а для $m=0, 1, \dots, k-2$ произведение $x y_m$ есть многочлен степени, меньшей k , и, следовательно, ортогонально y_k . Таким образом, $\gamma_0 = \gamma_1 = \dots = \gamma_{k-2} = 0$.

Для $m=k-1$ равенство (17.10-3) принимает вид

$$\begin{aligned} - \int_a^b \rho y_k (x y_{k-1}) dx &= - \int_a^b \rho y_k \left(\frac{\alpha_{k-1}}{\alpha_k} y_k + c'_{k-1} y_{k-1} + \dots \right) dx = \\ &= - \frac{\alpha_{k-1}}{\alpha_k} \lambda_k = \gamma_{k-1} \lambda_{k-1}; \end{aligned}$$

используя (17.10-1), получаем

$$c_k = \gamma_{k-1} = - \frac{\alpha_{k-1} \lambda_k}{a_k \lambda_{k-1}} \neq 0.$$

Для $m=k$ (17.10-3) принимает вид (используя (17.10-1))

$$\int_a^b \rho (a_k y_{k+1} - x y_k) y_k dx = - \int_a^b \rho x y_k^2 dx = \gamma_k = b_k.$$

Таким образом, равенство (17.10-2) может быть записано в виде

$$\frac{\alpha_k}{\alpha_{k+1}} y_{k+1} - x y_k = \gamma_k y_k + \gamma_{k-1} y_{k-1}$$

или

$$y_{k+1} = \frac{\alpha_{k+1}(\gamma_k + x)}{\alpha_k} y_k - \frac{\alpha_{k-1} \alpha_{k+1}}{\alpha_k^2} \cdot \frac{\lambda_k}{\lambda_{k-1}} y_{k-1}. \quad (17.10-4)$$

Из этой формулы следует несколько важных результатов. Покажем, прежде всего, что k нулей многочлена $y_k(x)$ разделены $k-1$ нулями многочлена $y_{k-1}(x)$ при условии, что $\rho(x) \geq 0$. Доказательство проведем по индукции. На первом шаге имеем $y_0(x) = \alpha_0 > 0$ и $y_1(x) = \alpha_1(x + \gamma_0)$. Как известно, $y_1(x)$ имеет действительный корень в интервале интегрирования, так как

$$\int_a^b \rho(x) y_0(x) y_1(x) dx = 0.$$

Предположим, что нули $y_k(x)$ разделены нулями $y_{k-1}(x)$. В нулях $y_k(x_i)$ (17.10-4) принимает вид

$$y_{k+1}(x_i) = - \frac{\alpha_{k-1} \alpha_{k+1}}{\alpha_k^2} \cdot \frac{\lambda_k}{\lambda_{k-1}} y_{k-1}(x_i). \quad (17.10-5)$$

В конце интервала $x=b$ как $y_{k+1}(b)$, так и $y_{k-1}(b)$ положительны, поскольку старшие коэффициенты выбраны положительными $\alpha_i > 0$,

а все нули функции лежат внутри интервала. В наибольшем нуле $y_k(x_i) = 0$, $y_{k-1}(x_i)$ положительно; следовательно, $y_{k+1}(x_i)$ отрицательно в силу (17.10-5). В следующем по величине нуле y_{k-1} по предположению индукции имеет другой знак, следовательно, $y_{k+1}(x)$ положительно, и т. д. По мере того как мы переходим от одного нуля y_k к следующему нулю, многочлен y_{k+1} тоже меняет знак. Таким образом, мы показали, что нули $y_k(x)$ разделяют нули $y_{k+1}(x)$, за исключением наименьшего нуля $y_{k+1}(x)$. Но так как мы показали, что $y_{k+1}(x)$ имеет $k+1$ различных действительных корней на (a, b) , последний нуль $y_{k+1}(x)$ меньше всех нулей $y_k(x)$, и доказательство индукции закончено.

Есть еще одно важное следствие трехчленных рекуррентных соотношений (17.10-1) или (17.10-2). Как только известны коэффициенты рекуррентного соотношения как функции k , а также $y_0(x)$ и $y_1(x)$, то можно последовательно вычислять $y_k(x)$ быстрее, чем громоздким методом Шмидта (см. § 18.2).

§ 17.11. Классические ортогональные многочлены

Существуют три подробно исследованных множества ортогональных многочленов: многочлены Лежандра, $P_n(x)$

$$\int_{-1}^1 P_m(x) P_n(x) dx = \begin{cases} 0, & m \neq n, \\ \frac{2}{2n+1}, & m = n, \end{cases}$$

многочлены Лагерра, $L_n(x)$

$$\int_0^{\infty} e^{-x} L_m(x) L_n(x) dx = \begin{cases} 0, & m \neq n, \\ \frac{1}{n!}, & m = n, \end{cases}$$

и многочлены Эрмита, $H_n(x)$

$$\int_{-\infty}^{\infty} e^{-x^2} H_m(x) H_n(x) dx = \begin{cases} 0, & m \neq n, \\ 2^n n! \sqrt{\pi}, & m = n. \end{cases}$$

Соответствующие им трехчленные рекуррентные соотношения (см. (17.10-4)) имеют вид

$$\begin{aligned} (n+1)P_{n+1}(x) - (2n+1)xP_n(x) + nP_{n-1}(x) &= 0, \\ (n+1)L_{n+1}(x) - (2n+1-x)L_n(x) + n^2L_{n-1}(x) &= 0, \quad (n \geq 1) \\ H_{n+1}(x) - 2xH_n(x) + 2nH_{n-1}(x) &= 0, \end{aligned}$$

Много других соотношений, касающихся этих функций, можно найти в обычных работах. Нули многочленов Лежандра, например, оказываются узлами гауссовых квадратурных формул (гл. 10).

Соответственно этим ортогональным функциям, определенным для интегрирования, существуют ортогональные функции для дискретных множеств. Из них наиболее распространено множество функций, соответствующее полиномам Лежандра. Из соображений удобства иногда выбирают интервал $-1 \leq x_i \leq 1$, а иногда от 0 до 1 при равноотстоящих узлах x_i , и читатель должен внимательно следить за тем, какое из этих множеств использовано в таблице.

Упражнение 17.11-1. Если $P_0 = 1$, $P_1 = x$, показать, что для $-1 \leq x \leq 1$

$$P_2 = \frac{1}{2}(3x^2 - 1), \quad P_3 = \frac{1}{2}(5x^3 - 3x), \quad P_4 = \frac{1}{8}(35x^4 - 30x^2 + 3).$$

§ 17.12. Сравнение метода наименьших квадратов и разложения в степенные ряды

Разложение в степенной ряд функции $y(x)$

$$y(x) = \sum_{n=0}^{\infty} \frac{y^{(n)}(0) x^n}{n!},$$

будучи усеченным до

$$y_N(x) = \sum_{n=0}^N \frac{y^{(n)}(0) x^n}{n!},$$

дает очень хорошее приближение вблизи $x=0$, но когда x возрастает, это приближение имеет тенденцию ухудшаться. С другой стороны, метод наименьших квадратов пытается найти более или менее равномерное (в зависимости от весовой функции $\rho(x)$) приближение на интервале. И в этом состоит одно из основных различий: степенные ряды приближают в точке, в то время как наименьшие квадраты — на интервале.

Когда дело доходит до непосредственных вычислений, то не всегда возможно точно вычислить производную в точке, используя узлы, разбросанные по интервалу; в этом случае вместо усеченного ряда часто используют точно приближающий интерполяционный полином.

В случае наименьших квадратов часто возникает возможность выбора. Можно представить себе, что имеются непрерывные ортогональные полиномы, коэффициенты Фурье (интегралы) по которым можно вычислить каким-либо численным методом. В другом случае можно начать с дискретных узлов и вести вычисления, используя дискретное ортогональное множество функций. Трудно сказать, какой из методов дает лучшие результаты в большинстве случаев. По всей вероятности, выбор зависит от обстоятельств, включая предпочтения лица, пользующегося результатами.

Нужно заметить, что при работе с рядами Фурье возможны два подхода, приводящих к совершенно одинаковым вычислениям. Эти подходы напоминают различные возможные взгляды на преобразование системы координат в аналитической геометрии, о которых шла речь в § 1.1. Можно пользоваться любым из них, в зависимости от удобства, при условии, что мы всегда будем помнить о содержательной стороне дела.

§ 17.13. Метод наименьших квадратов с ограничениями; продолжение примера из § 1.9

В примере, приведенном в § 1.9, мы строили интерполяционный многочлен по 11 точкам на отрезке $[0, 1]$ с шагом $h=0,1$, используя интерполяционную формулу Ньютона. Благодаря результатам этих вычислений было создано новое лабораторное оборудование, позволившее собрать более точные и более обширные данные.

Визуальное исследование данных показало, что результаты аппроксимируются многочленом шестой степени, а график значений вблизи точки $x=1$ имел наклон около 6, что, казалось, подтверждало эти исследования. «На радостях» был проведен многочлен методом наименьших квадратов, но результаты оказались никуда не годными. В точке $x=0$ многочлен имел положительное значение, в то время как эксперимент давал чистый нуль, после чего многочлен становился отрицательным, что физически невозможно.

Наиболее простым оказалось отбросить свободный член многочлена. Член, содержащий x , отбросить было нельзя, поскольку из обсуждения с заказчиком выяснилось, что кривая не должна касаться оси x в точке $x=0$. Вычисления, выполненные с учетом этих замечаний, дали вполне приемлемый многочлен, а остальная часть вычислений была столь же простой, как и раньше.

Другой очевидный способ решения задачи аппроксимации этих данных в точке $x=0$ состоит в том, чтобы придать значению в этой точке большой вес, скажем 1000, а остальным значениям данных придать веса, равные 1.

Существует много вариаций на тему наложения дополнительных условий на метод наименьших квадратов. Например, предположим, что в точках a и b решение должно принимать два заданных значения и аппроксимация ведется ортогональными полиномами. В этом случае можно использовать классический метод множителей Лагранжа. Если бы значения в точках a и b были равны 0, то многочлен $P_n(x)$ можно было бы построить, используя весовую функцию $(x-a)^2(x-b)^2$. Тогда

$$Q_n(x) = (x-a)(x-b)P_n(x)$$

будут ортогональными многочленами, проходящими через нуль в точках $x=a$ и $x=b$.

Упражнение 17.13-1. Показать, что в методе ортогональных полиномов из начальных данных можно так вычесть линейную функцию, что условия $f(a) = A$ и $f(b) = B$ сведутся к условию $\zeta(a) = \zeta(b) = 0$.

§ 17.14. Последние замечания о методе наименьших квадратов

Получив некоторое представление о проведении многочленов методом наименьших квадратов, читатель, возможно, ожидает, что автор перейдет к рассмотрению задач интегрирования и решения дифференциальных уравнений. Несмотря на реальность этого, метод наименьших квадратов редко используется в качестве основы при интегрировании. Интегрирование является процессом сглаживания. Если дифференциальное уравнение имеет «шумовой член», то обычно его сглаживают, и после этого для интегрирования уравнений используют точно аппроксимирующий полином.

Другая идея, которая может возникнуть у читателя, но которая еще недостаточно изучена, состоит в следующем: в методе точной аппроксимации мы полагали $E_0, E_1, \dots, E_{m-1} = 0$, а остальные E_m, E_{m+1}, \dots оставляли теми, какие они есть. Предположим, что мы пытаемся минимизировать выражение

$$m(a_0, a_1, \dots) = \sum_{k=0}^{\infty} a_k E_k^2.$$

В процессе точной аппроксимации $a_m = a_{m+1} = \dots = 0$, но можно слегка изменить a_k и таким образом повысить точность аппроксимации для более высоких степеней за счет точности более низких. Это можно рассматривать как предложение аппроксимировать в пространстве ошибок, а не в пространстве функций. Подобная теория остаточного члена еще в настоящее время не разработана.

Примером незначительного уменьшения точности аппроксимации для некоторого значения k , приводящим к большему повышению точности при более высоком значении k , является идея «отбрасывания», кратко рассмотренная в § 9.5 в связи с интерполяционной формулой Эверетта. В данном случае мы поступились небольшой точностью в x^2 с тем, чтобы намного улучшить аппроксимацию x^3 и x^4 .

ГЛАВА 18

МЕТОД НАИМЕНЬШИХ КВАДРАТОВ. ПРАКТИКА

§ 18.1. Общие замечания о многочленном случае

Метод наименьших квадратов используется главным образом в тех ситуациях, где надо определить коэффициенты, входящие линейно, особенно как коэффициенты многочлена. Рассмотрим сначала именно такой случай, а затем перейдем к другим.

Как было указано в предыдущей главе, определитель нормального уравнения (17.5-4) обычно бывает очень мал; поэтому решение относительно коэффициентов должно быть довольно неопределенным. Необходимо, однако, различать две вещи: точность коэффициентов и малость суммы квадратов ошибок. Если упомянутый определитель мал, то коэффициенты будут найдены плохо; тем не менее сумма квадратов ошибок может быть близка к минимуму. Вообще говоря, когда число определяемых коэффициентов не превосходит пяти-шести, прямое решение нормального уравнения обычно приемлемо; но при большем их числе скорее всего встретятся трудности.

По теории расширение системы ортогональных многочленов заменяет прямое решение. Однако опыт показывает, что если проводить ортогонализацию при помощи процесса Шмидта, то возникнут те же трудности, хотя и в другом виде. В процессе Шмидта m -й многочлен строится с помощью первых разностей всех компонент вектора x^m , т. е. $(x_1^m, x_2^m, \dots, x_N^m)$, который лежит в направлении ранее определенных многочленов. Результат, как правило, бывает мал, если только m недостаточно велико, и в качестве последнего шага его еще требуется нормировать. Такое нормирование увеличивает ошибку, потому что нормирующий множитель обычно имеет величину, много большую единицы.

Грубо говоря, затруднение можно выразить замечанием, что для больших n вектор x^n направлен почти туда же, куда и x^{n-1} , x^{n-2} и т. д. Таким образом, уравнения, ведущие к определителю Гильберта, почти линейно зависимы; отсюда этот определитель мал.

§ 18.2. Трехчленное рекуррентное соотношение

Если все-таки пытаться использовать ортогональные многочлены, то, по-видимому, можно избежать затруднений, упомянутых в предыдущем параграфе, используя для образования ортогональных многочленов трехчленное рекуррентное соотношение (17.10-2).

Запишем (17.10-2) в виде

$$\left. \begin{aligned} p_0(x) &= 1, \\ p_1(x) &= xp_0(x) - \alpha_1 p_0(x), \\ p_{k+1}(x) &= xp_k(x) - \alpha_{k+1} p_k(x) - \beta_{k+1} p_{k-1}(x), \quad k \geq 1, \end{aligned} \right\} \quad (18.2-1)$$

где α_{k+1} и β_{k+1} подлежат определению. Найдем сначала α_1 . Известно что

$$\int_a^b \rho(x) p_0(x) \gamma_1(x) dx = 0.$$

Отсюда по (18.2-1)

$$\int_a^b \rho(x) x dx = \alpha_1 \int_a^b \rho(x) dx.$$

Предположим теперь, что известны $p_0(x)$, $p_1(x)$, ..., $p_k(x)$ и что они взаимно ортогональны. Требуется вычислить следующий многочлен системы $p_{k+1}(x)$. Потребуем сначала, чтобы

$$\int \rho(x) p_{k+1}(x) p_k(x) dx = 0, \quad \int \rho(x) p_{k+1}(x) p_{k-1}(x) dx = 0.$$

Этого достаточно, чтобы определить α_{k+1} и β_{k+1} . Используя определение $p_{k+1}(x)$ (равенства (18.2-1)), получаем

$$\begin{aligned} \int \rho(x) x p_k^2(x) dx &= \alpha_{k+1} \int \rho(x) p_k^2(x) dx + \beta_{k+1} \int \rho(x) p_k(x) p_{k-1}(x) dx, \\ \int \rho(x) x p_k(x) p_{k-1}(x) dx &= \\ &= \alpha_{k+1} \int \rho(x) p_k(x) p_{k-1}(x) dx + \beta_{k+1} \int \rho(x) p_{k-1}^2(x) dx. \end{aligned}$$

Поскольку $p_k(x)$ и $p_{k-1}(x)$ ортогональны, имеем

$$\alpha_{k+1} = \frac{\int \rho(x) x p_k^2(x) dx}{\int \rho(x) p_k^2(x) dx}, \quad \beta_{k+1} = \frac{\int \rho(x) x p_k(x) p_{k-1}(x) dx}{\int \rho(x) p_{k-1}^2(x) dx}.$$

Знаменатель выражения для β_{k+1} был уже вычислен на предыдущем шаге, когда определялось $\alpha_k(x)$; таким образом, на каждом шаге надо вычислить три интеграла.

Ортогональность полученного многочлена $p_{k+1}(x)$ всем $p_i(x)$ ($i < k-1$) следует из равенства (18.2-1)

$$p_{k+1}(x) = x p_k(x) - \alpha_{k+1} p_k(x) - \beta_{k+1} p_{k-1}(x),$$

поскольку, умножая его на $\rho(x) p_i(x)$ и интегрируя, получаем

$$\begin{aligned} \int \rho(x) p_k(x) [x p_i(x)] dx - \alpha_{k+1} \int \rho(x) p_k(x) p_i(x) dx - \\ - \beta_{k+1} \int \rho(x) p_{k-1}(x) p_i(x) dx = 0. \end{aligned}$$

Два последних интеграла равны нулю в силу ортогональности, а в первом $x p_i(x)$ есть многочлен степени меньшей, чем k ; поэтому этот интеграл тоже равен нулю.

Если попытаться построить многочлены на дискретном множестве точек x_j ($j = 1, \dots, N$), то интегралы заменятся суммами. Если пытаться строить более N многочленов (т. е. сверх $p_{N-1}(x)$), то уравнение, определяющее α_N , выродится, как и должно быть.

Упражнение 18.2-1. Ортогонализировать систему многочленов $1, x, x^2, x^3$ на интервале $0 \leq x \leq 1$, используя трехчленный рекуррентный метод.

§ 18.3. Построение квазиортогональных многочленов

Прямой подход приводит к нормальной системе уравнений, у которой есть ненулевые члены вне главной диагонали. Использование ортогональных многочленов приводит к уравнениям, у которых все члены вне главной диагонали равны нулю; поэтому решение системы тривиально. Если строить ортогональные многочлены процессом Шмидта (или при помощи трехчленного рекуррентного соотношения), но определить их неточно, то у окончательных уравнений, которые придется решать, будут большие члены на главной диагонали, тогда как члены вне диагонали пропорциональны отклонению от ортогональности. Если оно невелико, то система уравнений совсем легко решается. Это наводит на мысль попробовать построить квазиортогональные многочлены, не делая больших вычислений. Насколько мы их сделаем ортогональными, настолько члены вне главной диагонали системы нормальных уравнений будут нулями, и наоборот, насколько мы ошибемся в ортогональности, настолько они будут далеки от нуля.

Размышления на эту тему приводят к мысли, что основное свойство, которое надо постараться сохранить, есть свойство чередования корней. Выберем первый многочлен равным единице, $p_0(x) = 1$. Вторым выберем прямую подходящего наклона и, кроме того, с нулем, близким к середине нашего множества точек. Далее выбираем параболу, нули которой лежат по разные стороны от нуля прямой, и т. д., на каждом шагу выбирая нули следующего многочлена так, чтобы они разделялись предшествующим многочленом, а также оставляли место для следующих многочленов. Пока мы не пытаемся построить многочлен степени, близкой к степени многочлена, проходящего через все наши точки (а это бывает редко, если бывает вообще), у нас есть достаточно свободы, чтобы выбрать корни в удобных местах, сделать удобными коэффициенты многочленов и тем самым облегчить бремя промежуточных вычислений.

Опыт работы по этому методу показывает, что он весьма эффективен; обычно недиагональные члены системы уравнений очень малы и уравнения легко решаются.

Упражнение 18.3-1. Рассмотрите квазиортогональные многочлены

$$(-1 \leq x \leq 1), \quad p = 1; \quad p_0 = 1; \quad p_1 = x; \quad p_2 = x^2 - \frac{1}{4}; \quad p_3 = x \left(x^2 - \frac{9}{16} \right).$$

§ 18.4. Немногочленный случай

Иногда бывает дана система функций $f_i(x)$ и требуется приблизить по методу наименьших квадратов некоторые данные функцией $f(x)$ вида

$$f(x) = a_1 f_1(x) + a_2 f_2(x) + \dots + a_n f_n(x).$$

Действуем как и раньше. Записываем ошибку ε_i в точке x_i и образуем сумму квадратов. Затем дифференцируем по коэффициентам (которые являются переменными задачи). Получившиеся нормальные уравнения формально подобны уравнениям в многочленном случае, за исключением того, что теперь определяются

$$S_{k,j} = S_{j,k} = \sum_i f_j(x_i) f_k(x_i); \quad T_k = \sum_i f(x_i) f_k(x_i).$$

Если функции $f_i(x)$ сильно линейно независимы, т. е. ни одна из них с единичным коэффициентом не может быть выражена как линейная комбинация остальных плюс малая поправка, то можно ожидать, что удастся решить нормальные уравнения без больших затруднений. Если они близки к линейно зависимым, то можно использовать похожую схему ортогонализации, хотя здесь, вообще говоря, не будет никакого трехчленного рекуррентного соотношения и квазиортогональный метод может оказаться трудно изобразимым, если у $f_i(x)$ много нулей.

§ 18.5. Нелинейные параметры

Часто случается, что теория дает вид формулы, которой должны удовлетворять некоторые данные, и эти данные должны использоваться для определения коэффициентов формулы, применяя критерий наименьших квадратов. Например, предположим, что данная формула такова:

$$y(x) = a + be^{cx}, \quad (18.5-1)$$

и следует определить a , b , c так, чтобы для данных (x_i, y_i) ($i = 1, \dots, N$)

$$\sum_{i=1}^N [y_i - (a + be^{cx_i})]^2 = \min = m. \quad (18.5-2)$$

Если продифференцировать по a , b , c соответственно и приравнять производные нулю, получатся уравнения, которые трудно решить.

Предположим, что данные нанесены на график и примерно определено, что $c = c_1$. Теперь можно определить a и b обычным способом. Затем можно вычислить $m(c_1)$ из (18.5-2) для этого значения c .

Далее, попробуем какое-нибудь другое возможное значение $c = c_2$ и снова вычислим $m = m(c_2)$. Исследование этих двух значений

$$\begin{array}{c|c} c & m(c) \\ \hline c_1 & m(c_1) \\ c_2 & m(c_2) \end{array}$$

подскажет (по линейности) новое значение c , скажем c_3 .

Таким способом можно приблизиться к минимальному значению $m(c)$. Детали стратегии поиска будут обсуждены позже. Очевидно, однако, что мы проделаем много больше вычислений по сравнению с тем случаем, когда все определяемые параметры входят линейно. Чем больше параметров входит нелинейно, тем больше вычислений требуется, чтобы найти приближение способом наименьших квадратов. Опыт показывает, что, когда число нелинейных параметров достигает четырех или пяти, процесс может быть крайне мучительным и медленным.

ГЛАВА 19

МНОГОЧЛЕНЫ ЧЕБЫШЕВА

§ 19.1. Введение

Ряды Фурье имеют много замечательных свойств. Отметим следующие из них:

1. Каждая функция является равноколеблющейся, т. е. чередующиеся максимумы и минимумы одинаковы.

2. Одни и те же аналитические выражения (синусы и косинусы) ортогональны как на непрерывном, так и на дискретном множестве равноотстоящих точек (гл. 6).

Чтобы увидеть, насколько замечательно второе свойство, рассмотрим многочлены Лежандра $P_n(x)$. Если первые m из них, $m=0, 1, \dots, n-1$ должны быть ортогональны на дискретном множестве точек, то чтобы определить положение n узловых точек (разрешая любое их расположение), понадобятся $\frac{n(n-1)}{2}$ уравнений

$$\sum_{i=1}^n P_j(x_i) P_k(x_i) = 0, \quad j \neq k.$$

Нежелательность этого уже для не слишком больших n очевидна. Это не значит, конечно, что нет системы многочленов, ортогональных на данном множестве, но члены этой системы зависят, вообще говоря, от n — числа используемых точек. Когда n стремится к бесконечности, члены системы, ортогональной на дискретном множестве точек, стремятся к соответствующим членам системы, ортогональной на непрерывном интервале.

В классе ортогональных функций подкласс ортогональных многочленов имеет ряд особых свойств:

1. Они удовлетворяют трехчленному рекуррентному соотношению.
2. Их легко вычислять и превращать в степенные ряды.
3. Их нули разделяют друг друга.

Многочлены Чебышева обладают всеми свойствами как рядов Фурье, так и ортогональных многочленов, они и являются, в сущности, функциями Фурье $\cos n\theta$, замаскированными простым преобразованием переменной

$$\theta = \arccos x. \quad (19.1-1)$$

Таким образом, многочлены Чебышева естественным образом играют уникальную роль среди ортогональных функций. Выражение Фурье для ортогональности становится таким:

$$\int_0^\pi \cos m\theta \cos n\theta d\theta = \begin{cases} 0 & (m \neq n) \\ \frac{\pi}{2} & (m = n \neq 0) \\ \pi & (m = n = 0) \end{cases} = \int_{-1}^1 T_m(x) T_n(x) \frac{dx}{\sqrt{1-x^2}}; \quad (19.1-2)$$

$$\sum_{j=0}^{N-1} \cos m\theta_j \cos n\theta_j = \begin{cases} 0 & (m \neq n) \\ \frac{N}{2} & (m = n \neq 0) \\ N & (m = n = 0) \end{cases} = \sum_{j=0}^{N-1} T_m(x_j) T_n(x_j).$$

Здесь обозначено *)

$$T_n(x) = \cos(n \arccos x)$$

и использована ортогональность $\cos nx$ на интервале $(0 \leq x \leq \pi)$ (см. § 6.7).

Покажем, что $T_n(x)$ — многочлен. По теореме Муавра

$$\cos n\theta + i \sin n\theta = (\cos \theta + i \sin \theta)^n.$$

Разлагая бином, взяв действительные части с обеих сторон и заменив четные степени $\sin \theta$ из

$$(\sin^2 \theta)^k = (1 - \cos^2 \theta)^k,$$

видим, что $\cos n\theta$ есть многочлен степени n от $\cos \theta$. Но $\cos(\arccos x) = x$, отсюда $T_n(x) = \cos(n \arccos x)$ есть многочлен степени n от x

*) Обозначение $T_n(x)$ происходит от французского написания фамилии Чебышева (Tschebyscheff).

Преобразование (19.1-1) можно рассматривать как проекцию пересечений полукруга с множеством прямых, имеющих равные углы между собой [рис. (19.1-1)]. Таким образом, множество точек x_j , на котором система чебышевских многочленов $T_n(x)$ ортогональна, таково:

$$x_j = \cos \frac{\pi}{N} j$$

$$(j=0, 1, \dots, N-1). \quad (19.1-3)$$

Это неравномерное расположение, у которого x_j сгущаются к обоим концам интервала $(-1 \leq x \leq 1)$, компенсируется в непрерывном случае (см. (19.1-2)) весовой функцией

$\frac{1}{\sqrt{1-x^2}}$. Таким образом, почти выполняется свойство 2 рядов Фурье; пришлось только отказаться от равномерного расположения узловых точек. Поскольку многочлены $T_n(x)$ есть, по существу, $\cos n\theta$, то они тоже являются равноколеблющимися функциями, и поскольку они многочлены, они обладают всеми свойствами ортогональных многочленов.

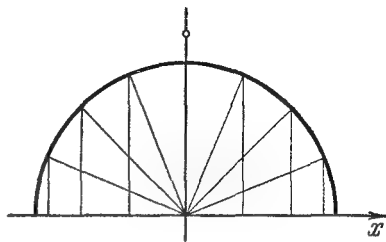


Рис. 19.1-1. Чебышевские узловые точки.

§ 19.2. Некоторые тождества

Многие свойства многочленов Чебышева следуют из соответствующих тождеств для тригонометрических функций. Например, тождество

$$\cos(n+1)\theta + \cos(n-1)\theta = 2 \cos \theta \cos n\theta$$

становится, согласно (19.1-3), равенством

$$T_{n+1}(x) + T_{n-1}(x) = 2xT_n(x) \quad (n \geq 1), \quad (19.2-1)$$

которое является трехчленным рекуррентным соотношением, соответствующим (17.10-1).

Из тождества

$$\cos(m+n)\theta + \cos(m-n)\theta = 2 \cos m\theta \cos n\theta,$$

получаем

$$T_{m+n}(x) + T_{m-n}(x) = 2T_m(x)T_n(x).$$

Пологая $m=n$, имеем равенство

$$T_{2n}(x) = 2T_n^2(x) - 1, \quad (19.2-2)$$

которое иногда полезно для получения одного многочлена высокого порядка.

Из (19.2-1) легко вывести, что старший член многочлена $T_n(x)$ имеет вид

$$\begin{aligned} T_n(x) &= 2^{n-1}x^n + \dots \quad (n \geq 1), \\ T_0(x) &= 1 \quad (n = 0) \end{aligned} \quad (19.2-3)$$

и что $T_n(x)$ — многочлен четной или нечетной степени соответственно тому, четно или нечетно n . Поскольку

$$T_{n+1} = \cos [(n+1) \arccos x]$$

имеет производную

$$\frac{1}{n+1} \frac{dT_{n+1}}{dx} = \frac{-\sin [(n+1) \arccos x]}{-\sqrt{1-x^2}},$$

то

$$\begin{aligned} \frac{1}{n+1} \frac{dT_{n+1}}{dx} - \frac{1}{n-1} \frac{dT_{n-1}}{dx} &= \frac{T'_{n+1}}{n+1} - \frac{T'_{n-1}}{n-1} = \\ &= \frac{\sin (n+1)\theta - \sin (n-1)\theta}{\sin \theta} = \frac{2 \cos n\theta \sin \theta}{\sin \theta} = 2T_n. \end{aligned} \quad (19.2-4)$$

Если переписать это равенство в виде

$$\frac{T'_n}{n} = 2T_{n-1} + \frac{T'_{n-2}}{n-2} \quad (n > 2),$$

то приходим к выражениям:

$$\begin{aligned} \frac{T'_{2n}}{2n} &= 2(T_{2n-1} + T_{2n-3} + \dots + T_1), \\ \frac{T'_{2n+1}}{2n+1} &= 2(T_{2n} + T_{2n-2} + T_{2n-4} + \dots + T_0) + 1 \end{aligned} \quad (19.2-5)$$

для четного и нечетного случая соответственно.

Упражнения

19.2-1. Вывести равенство (19.2-5).

19.2-2. Используя (19.2-1), доказать

$$x^k T_n(x) = \frac{T_{n+k} + C(k, 2) T_{n+k-2} + C(k, 2) T_{n+k-4} + \dots + T_{n-k}}{2^k},$$

19.2-3. Доказать

$$\frac{T_{2k+1}(x)}{x} = 2T_{2k} - 2T_{2k-2} + 2T_{2k-4} - \dots \pm T_0.$$

§ 19.3. Критерий Чебышева

Чебышев показал, что из всех многочленов $P_n(x)$ степени n со старшим коэффициентом 1 у многочлена $\frac{T_n(x)}{2^{n-1}}$ точная верхняя грань абсолютных значений на интервале $(-1 \leq x \leq 1)$ наименьшая. По-

сколько верхняя грань $|T_n(x)|$ равна 1, указанная верхняя грань равна $\frac{1}{2^{n-1}}$.

Доказательство этого замечательного свойства вытекает из рассмотрения разности

$$\varphi_{n-1}(x) = \frac{T_n(x)}{2^{n-1}} - P_n(x),$$

которая есть многочлен степени $(n-1)$ (так как члены x^n уничтожаются [см. (19.2-3)]). Из того факта, что $T_n(x)$ есть $\cos n\theta$, видно, что $T_n(x)$ в интервале $(-1 \leq x \leq 1)$ принимает свое экстремальное значение $n+1$ раз, по очереди положительным и отрицательным. Если экстремальное значение у $P_n(x)$ меньше, чем у $\varphi_{n-1}(x)$, то в этих $n+1$ экстремальных точках $\varphi_{n-1}(x)$ по очереди положительно и отрицательно; следовательно, у нее должны быть n действительных корней между этими точками. Так как $\varphi_{n-1}(x)$ имеет степень $n-1$, то можно заключить, что $\varphi_{n-1}(x) \equiv 0$ и $P_n(x) = \frac{T_n(x)}{2^{n-1}}$.

Это свойство представляет большой интерес в численном анализе. Если какая-либо ошибка может быть выражена многочленом Чебышева степени n , то любое другое выражение для ошибки в виде многочлена степени n , имеющего тот же самый старший коэффициент, будет иметь на интервале $(-1 \leq x \leq 1)$ большую максимальную ошибку, чем чебышевское. В соответствии с этим, «чебышевским приближением» называют такое*), при котором стремятся свести к минимуму максимум ошибки. Иногда это называют «принципом минимакса». Приближение в смысле наименьших квадратов уменьшает среднюю квадратичную ошибку, но при этом допускает отдельные большие ошибки; чебышевское — уменьшает экстремальную ошибку, допуская большое среднеквадратичное отклонение.

В качестве простой иллюстрации рассмотрим задачу интерполяции многочленом степени n на интервале $(-1 \leq x \leq 1)$. Остаточный член (8.6-1) имеет вид

$$\frac{(x-x_1)(x-x_2) \dots (x-x_{n+1}) y^{(n+1)}(\bar{x})}{(n+1)!}.$$

Если мы хотим минимизировать максимальное отклонение за счет множителя $(x-x_1)(x-x_2) \dots (x-x_{n+1})$, то нужно выбрать узловые точки в нулях многочлена $T_{n+1}(x)$. Тем самым множитель, который легко контролировать, будет равноколеблющимся многочленом с наименьшим максимальным отклонением.

*) Чебышевское приближение не следует путать с чебышевским интегрированием; существует довольно много разнообразных идей, носящих его имя.

§ 19.4. Экономизация

Другой простой, но очень важный пример использования чебышевских многочленов — процесс «экономизации»*), принадлежащий в основном Ланцошу. Пусть дан отрезок степенного ряда функции

$$f(x) = a_0 + a_1x + a_2x^2 + \dots + a_Nx^N \quad (19.4-1)$$

в интервале $(-1 \leq x \leq 1)$. Для степенного ряда ошибка обычно бывает велика на концах интервала и мала в середине. Процесс экономизации начинается с того, что мы, используя таблицу

$$\left. \begin{aligned} 1 &= T_0, & x^4 &= \frac{1}{8}(3T_0 + 4T_2 + T_4), \\ x &= T_1, & x^5 &= \frac{1}{16}(10T_1 + 5T_3 + T_5), \\ x^2 &= \frac{1}{2}(T_0 + T_2), & \dots & \\ x^3 &= \frac{1}{4}(3T_1 + T_3), & \dots & \end{aligned} \right\} \quad (19.4-2)$$

превращаем степенной ряд в разложение по многочленам Чебышева

$$f(x) = b_0 + b_1T_1(x) + b_2T_2(x) + \dots + b_NT_N(x). \quad (19.4-3)$$

Это — разложение по ортогональным многочленам. Для широкого класса функций разложение по чебышевским многочленам сходится много быстрее, чем по любой другой системе ортогональных многочленов (обоснование этого см. в § 19.5). Таким образом, мы надеемся, что b_k в формуле (19.4-3) убывают много быстрее, чем a_k в (19.4-1).

Для иллюстрации рассмотрим пример, достаточно простой для того, чтобы его легко можно было просчитать:

$$y = \ln(1+x) \approx x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5}.$$

Используя (19.4-2), получаем

$$\begin{aligned} y &= T_1 - \frac{1}{4}(T_0 + T_2) + \frac{1}{12}(3T_1 + T_3) - \frac{1}{32}(3T_0 + 4T_2 + T_4) + \\ &+ \frac{1}{80}(10T_1 + 5T_3 + T_5) = -\frac{11}{32}T_0 + \frac{11}{8}T_1(x) - \frac{3}{8}T_2(x) + \\ &+ \frac{7}{48}T_3(x) - \frac{1}{32}T_4(x) + \frac{1}{80}T_5(x). \end{aligned} \quad (19.4-4)$$

В интервале $(0 \leq x \leq 1)$ отбрасывание последнего члена степенного ряда (19.4-1) дает изменение на $1/5$ (для $x=1$), в то же время

*) В русском переводе книги Ланцоша (Практические методы прикладного анализа, Физматгиз, М., 1961) этот метод назван «телескопическим сдвигом путем последовательного сокращения». (Прим. ред.)

отбрасывание трех последних членов (19.4-4) даст изменение меньше чем на

$$\frac{7}{48} + \frac{1}{32} + \frac{1}{80} = \frac{19}{15 \cdot 32} < \frac{1}{5.2},$$

так как $|T_k(x)| \leq 1$ на интервале $(-1 \leq x \leq 1)$. Таким образом, можно представить

$$y = \ln(1+x) \approx -\frac{11}{32} T_0 + \frac{11}{8} T_1(x) - \frac{3}{8} T_3(x)$$

многочленом второй степени в чебышевской форме более точно, чем соответствующим отрезком степенного ряда. Чебышевское разложение можно снова превратить в многочлен по таблице

$$\left. \begin{array}{ll} T_0 = 1, & T_4 = 8x^4 - 8x^3 + 1, \\ T_1 = x, & T_5 = 16x^5 - 20x^3 + 5x, \\ T_2 = 2x^2 - 1, & \dots \dots \dots \\ T_3 = 4x^3 - 3x, & \end{array} \right\} \quad (19.4-5)$$

Итак,

$$\begin{aligned} y = \ln(1+x) &\approx -\frac{11}{32} + \frac{11}{8}x - \frac{3}{8}(2x^3 - 1) = \\ &= \frac{1}{32} + \frac{11}{8}x - \frac{3}{4}x^3 \quad (0 \leq x \leq 1). \end{aligned}$$

Вообще говоря, можно ожидать, что степенной ряд, состоящий из большого числа членов, превращенный в чебышевское разложение, дает приближение многочленом значительно меньшей степени, так как можно отбрасывать много последних членов чебышевского разложения без большого увеличения ошибки по сравнению с тем, что давал первоначальный отрезок степенного ряда.

§ 19.5. Механизация процесса экономизации

Описанный процесс экономизации использует таблицы (19.4-2) и (19.4-5). Первая из них — просто хорошо известное тригонометрическое тождество

$$\begin{aligned} (\cos \theta)^k &= \left(\frac{e^{i\theta} + e^{-i\theta}}{2} \right)^k = \frac{1}{2^{k-1}} \times \\ &\times \left[\frac{e^{ik\theta} + e^{-ik\theta}}{2} + C_k^1 \frac{e^{i(k-2)\theta} + e^{-i(k-2)\theta}}{2} + C_k^2 \frac{e^{i(k-4)\theta} + e^{-i(k-4)\theta}}{2} + \dots \right], \end{aligned}$$

где последний член

$$\begin{cases} C_k^m \cos \theta, & \text{если } k = 2m + 1, \\ \frac{1}{2} C_k^m, & \text{если } k = 2m. \end{cases}$$

Можно построить обе таблицы последовательно, используя равенство (19.2-1). Однако, вместо того чтобы использовать таблицы, по-видимому, лучше запрограммировать этот процесс простым способом. Один такой способ основан на подстановке $x = \cos \theta$ в первоначальный степенной ряд

$$y = \sum_{k=0}^N a_k x^k = \sum_{k=0}^N a_k \cos^k \theta = \\ = a_0 + \cos \theta \{a_1 + \cos \theta [a_2 + \dots + (a_{N-1} + \cos \theta \cdot a_N) \dots]\}.$$

Начиная с тривиального ряда Фурье в круглых скобках

$$a_{N-1} + a_N \cos \theta$$

и умножая на $\cos \theta$, получим

$$\frac{1}{2} a_N + a_{N-1} \cos \theta + \frac{1}{2} a_N \cos 2\theta,$$

т. е. снова ряд Фурье.

Вообще, если на k -м шаге был ряд Фурье

$$a_0^{(k)} + a_1^{(k)} \cos \theta + a_2^{(k)} \cos 2\theta + \dots + a_k^{(k)} \cos k\theta,$$

то, умножая на $\cos \theta$, получаем

$$\frac{a_0^{(k)}}{2} + \left(a_0^{(k)} + \frac{a_2^{(k)}}{2}\right) \cos \theta + \left(\frac{a_1^{(k)}}{2} + \frac{a_3^{(k)}}{2}\right) \cos 2\theta + \dots + \frac{a_k^{(k)}}{2} \cos (k+1)\theta,$$

т. е. снова ряд Фурье, но содержащий один член более высокого порядка.

Иными словами, нулевой коэффициент k -го шага входит на $(k+1)$ -м шаге как слагаемое в коэффициент при $\cos \theta$; каждый коэффициент (кроме нулевого), деленный пополам, входит как в следующий, так и в предыдущий член по отношению к тому месту, где этот коэффициент стоял на k -м шаге.

Этой простой процедуры достаточно, чтобы вычислить коэффициенты чебышевского разложения из разложения в степенной ряд. Отсюда видно, что

$$b_N = \frac{a_N}{2^{N-1}}.$$

Грубо говоря, если рассматривать коэффициенты как массы, то этот процесс обладает свойством сохранения масс.

Более детальное рассмотрение этого процесса показывает, что b_k для больших k стремятся к нулю быстрее, чем a_k . Хотя это и оправдывает утверждение, что чебышевское разложение обычно сходится быстрее, чем степенной ряд, но не доказывает его.

Весь процесс может быть выведен непосредственно из рекуррентного соотношения (19.2-1), записанного в виде

$$\begin{aligned} xT_n &= \frac{1}{2} T_{n+1} + \frac{1}{2} T_{n-1}, & n > 1, \\ xT_0 &= T_1, & n = 1, \end{aligned}$$

но наш вывод показывает также, как от разложения в степенной ряд перейти к соответствующему ряду Фурье.

Переходя теперь к процессу экономизации, рассмотрим коэффициенты b_k чебышевского разложения и отбросим все те суммы, которые меньше допустимой ошибки (вместе с ошибкой отрезка степенного ряда).

Чтобы снова получить многочлен, обратим описанный выше процесс. Это обращение возможно, так как старший член возникает только в одном месте предыдущей строки. Отправляясь от старшего члена $(k+1)$ -й строки, мы можем вычислить старший член k -й строки. Заметим, что этот последний член влияет на второй с конца член

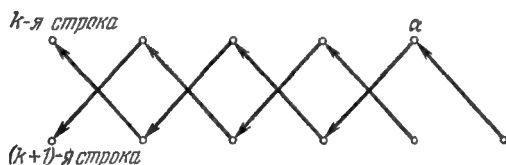


Рис. 19.5-1. Процесс перехода к многочленному виду.

$(k+1)$ -й строки. Возьмем затем соседний член с меньшим номером и продвинемся к началу строки (рис. 19.5-1). Последний шаг определит соответствующий коэффициент степенного ряда $a_0^{(k)}$. Как на прямом, так и на обратном шаге сумма коэффициентов сохраняется и это может служить контролем в конце, поскольку

$$\sum_{k=0}^N a_k = \sum_{k=0}^N b_k.$$

Последнее равенство легко получить, полагая $x=1$ ($T_k(x)=1$).

Итак, процесс экономизации легко механизировать без использования обширных таблиц.

Упражнение 19.5-1. Нарисуйте блок-схему процесса экономизации.

§ 19.6. Смещенные многочлены Чебышева

Часто удобно вместо интервала $-1 \leq x \leq 1$ использовать интервал $0 \leq x \leq 1$. Для этой цели употребляются смещенные многочлены Чебышева

$$T_n^*(x) = T_n(2x - 1). \quad (19.6-1)$$

Имеем

$$\begin{aligned}T_0^*(x) &= 1, \\T_1^*(x) &= 2x - 1, \\T_2^*(x) &= 8x^2 - 8x + 1, \\T_3^*(x) &= 32x^3 - 48x^2 + 18x - 1, \\T_4^*(x) &= 128x^4 - 256x^3 + 160x^2 - 32x + 1, \\&\dots\end{aligned}$$

и

$$\begin{aligned}1 &= T_0^*, \\x &= \frac{1}{2}(T_0^* + T_1^*), \\x^2 &= \frac{1}{8}(3T_0^* + 4T_1^* + T_2^*), \\x^3 &= \frac{1}{32}(10T_0^* + 15T_1^* + 6T_2^* + T_3^*), \\x^4 &= \frac{1}{128}(35T_0^* + 56T_1^* + 28T_2^* + 8T_3^* + T_4^*), \\&\dots\end{aligned}$$

Имеются более подробные таблицы *).

Для смещенных многочленов Чебышева справедливо рекуррентное соотношение

$$T_{n+1}^*(x) = (4x - 2) T_n^*(x) - T_{n-1}^*(x), \quad T_0^* = 1,$$

или

$$x T_n^*(x) = \frac{1}{4} T_{n+1}^*(x) + \frac{1}{2} T_n^*(x) + \frac{1}{4} T_{n-1}^*(x), \quad (19.6-2)$$

где

$$T_n^*(x) = \cos [n \arccos (2x - 1)] = T_n(2x - 1). \quad (19.6-3)$$

Упражнения

19.6-1. Примените смещенные многочлены T_n^* к примеру § 19.4. (Для этого придется продолжить таблицы $T_n^*(x)$ и x^n на одну строку.)

19.6-2. Проследите механизацию (процесса экономизации) для смещенных многочленов $T_n^*(x)$, как в § 19.4 [используйте равенство (19.6-2)].

19.6-3. Найдите для сдвинутых многочленов равенства, подобные (19.2-5).

§ 19.7 τ -процесс Ланцоша

Еще один прием использования многочленов Чебышева предложил Ланцош. В основе τ -процесса для решения линейных дифференциальных уравнений с полиномиальными коэффициентами лежит простая,

*) Tables of Chebyshev Polynomials $S_n(x)$ and $C_n(x)$, Natl. Bur. Standards (U. S.), Appl. Math. Series 9, 1952.

но важная идея. Обычно бывает трудно найти ошибку вычисленного решения задачи, но на обратный вопрос: «для какой близкой задачи только что вычисленный ответ был бы точным решением?» — ответить часто нетрудно. Это — один из ответов на четвертый основной вопрос § 7.1 «*что есть точность?*», и во многих случаях это — правильный ответ.

Чтобы проиллюстрировать τ-метод, предположим, что дано обыкновенное дифференциальное уравнение

$$y' + y = 0; \quad y(0) = 1,$$

и решение ищется в виде многочлена, т. е. мы надеемся оборвать степенной ряд, которым выражается решение. Очевидно, никакой многочлен не будет точно удовлетворять этому уравнению. Рассмотрим малое изменение правой части, прибавив к ней многочлен Чебышева, обозначая через τ величину максимального изменения

$$y' + y = \tau T_n^*(x).$$

Чтобы легко было проследить все вычисления, положим $n = 4$ и попробуем найти многочлен

$$y = a + bx + cx^2 + dx^3 + ex^4.$$

Приравнявая коэффициенты при одинаковых степенях x , получаем

$$\begin{aligned} b + a &= \tau, \\ 2c + b &= -32\tau, \\ 3d + c &= 160\tau, \\ 4e + d &= -256\tau, \\ e &= 128\tau. \end{aligned}$$

Начальное условие дает $a = 1$. Отсюда последовательно получаем

$$\begin{aligned} b &= \tau - 1, \\ c &= \frac{1 - 33\tau}{2}, \\ d &= \frac{1}{3} \left(160\tau + \frac{33\tau - 1}{2} \right) = \frac{353\tau - 1}{6}, \\ e &= \frac{1}{4} \left(-256\tau + \frac{1 - 353\tau}{6} \right) = \frac{1}{24} (1 - 1889\tau), \\ e &= 128\tau = \frac{1}{24} (1 - 1889\tau), \quad \tau = \frac{1}{4961}. \end{aligned}$$

Итак, ошибка, которая была внесена в первоначальное уравнение, равна

$$\left| \frac{T_4^*(x)}{4961} \right| \leq \frac{1}{4961},$$

и для нового уравнения имеем точное решение

$$y = 1 - \frac{4960}{4961}x + \frac{2464}{4961}x^2 - \frac{768}{4961}x^3 + \frac{128}{4961}x^4.$$

Сравним это с решением в виде степенного ряда

$$y = 1 - x + \frac{x^2}{2} - \frac{x^3}{6} + \frac{x^4}{24},$$

у которого максимальная ошибка примерно

$$\frac{1}{5!} = \frac{1}{120}.$$

В нашем случае легко оценить ошибку, происходящую от добавления чебышевского члена в дифференциальное уравнение,

$$y' + y = \tau T_4^*(x), \quad y(x) = e^{-x} \cdot 1 + \tau e^{-x} \int_0^x T_4^*(\theta) e^{\theta} d\theta.$$

Для $0 \leq x \leq 1$ справедливо неравенство

$$\begin{aligned} |y - e^{-x}| &\leq \left| \tau e^{-x} \int_0^x T_4^*(\theta) e^{\theta} d\theta \right| \leq \tau e^{-x} \int_0^x e^{\theta} d\theta = \tau(1 - e^{-x}) \leq \\ &\leq \frac{0,665}{4961} \approx 1,34 \times 10^{-4}; \end{aligned}$$

полученное значение много меньше $\frac{1}{5!}$.

Мы привели лишь простой пример τ -метода. В книге Ланцоша [23] содержится гораздо более широкое его изложение. Однако основная идея та же: мы слабо меняем условие задачи и получаем точное решение этой новой задачи. Поскольку в этом случае меняются условия первоначальной физической задачи, нам это обычно легче истолковать, чем какое-нибудь приближение, сделанное в ходе численного решения. τ -метод показывает, как важно аккуратно ответить на четвертый основной вопрос: «*что есть точность*»?

Упражнение 19.7-1. Применить τ -метод к уравнению

$$xy'' + y = 0; \quad y(0) = 0, \quad y'(0) = 1$$

для $(0 \leq x \leq 1)$, полагая $n = 4$.

§ 19.8. Видоизменение τ -метода

Предположим, что, вместо того чтобы действовать как Ланцош, мы пытаемся решить задачу в лоб и представляем решение дифференциального уравнения (с полиномиальными коэффициентами) в виде ряда по многочленам Чебышева

$$y(x) = \sum_{k=0}^{\infty} a_k T_k(x). \quad (19.8-1)$$

Тогда

$$y'(x) = \sum_{k=0}^{\infty} a_k T'_k(x),$$

и можно использовать (19.2-5),

$$y'(x) = \sum_{k=0}^{\infty} k a_k \left[2T_{k-1} + 2T_{k-3} + \dots + \binom{2T_1}{1} \right] * \quad (19.8-2)$$

и такие же выражения для старших производных, чтобы исключить все производные.

Затем используются равенства (19.2-1)

$$x T_n(x) = \frac{1}{2} [T_{n+1}(x) + T_{n-1}(x)] \quad (19.8-3)$$

и такие же выражения для $x^k T_n^*(x)$ (см. упражнение 19.2-2), чтобы исключить все степени x . Таким образом, мы приходим к ряду по чебышевским многочленам $T_n(x)$. Метод приравнивания коэффициентов при одинаковых степенях x в обеих частях тождества по x основывается на том, что разные степени x линейно независимы. Раз уж производные и члены с x исключены, т. е. имеется разложение лишь по T_n , то можно также приравнять коэффициенты при одинаковых чебышевских многочленах. Следовательно, можно действовать так, как если бы имели дело с разложением решения в степенной ряд. Оборвав ряд, получим ошибку, примерно равную первому отброшенному многочлену Чебышева.

Мы можем сделать этот оборванный ряд точным решением, прибавляя к правой части такие многочлены Чебышева, чтобы все члены уничтожились. Результат будет тот же, что и в описанном ранее τ -методе, но этот метод более гибок в вопросе о том, где оборвать решение.

Метод описан в терминах стандартных многочленов Чебышева; так же хорошо это можно сделать для смещенных многочленов $T_n^*(x)$.

Две формы одного и того же τ -метода приведены, чтобы подчеркнуть то обстоятельство, что если нужно получить ошибку в виде многочлена Чебышева, то, по-видимому, лучше начинать с представления решения в виде ряда по многочленам Чебышева, а не смешивать обыкновенные и чебышевские многочлены. Нужные тождества между многочленами легко найти.

*) Символ $\binom{2T_1}{1}$ в данном случае означает, что последним слагаемым может оказаться либо $2T_1$ (при четном k), либо 1 (при нечетном n). (Прим. ред.)

Упражнения

19.8-1. Изложите видоизмененный τ -метод, используя смещенные многочлены Чебышева $T_n^*(x)$.

19.8-2. Показать, что

$$\int T_n(x) dx = \frac{1}{2} \left[\frac{T_{n+1}(x)}{n+1} - \frac{T_{n-1}(x)}{n-1} \right] \quad (n > 1),$$

$$\int T_0(x) dx = T_1(x), \quad \int T_1(x) dx = \frac{1}{4} T_2(x),$$

т. е. что интегрирование может быть выполнено непосредственно в многочленах Чебышева.

19.8-3. Провести вычисления для примера $y' = y$ § 19.7 прямым τ -методом, как в § 19.8.

19.8-4. Применить прямой τ -метод к упражнению 19.7-1.

§ 19.9. Несколько замечаний о чебышевском приближении

Чебышевское приближение вызвало некоторый переполох среди тех, кто занимался вычислением таблиц, где важно лишь точно знать максимальную ошибку. Так, вместо того чтобы использовать для интерполяции многочлены, проходящие через заданные точки, было предложено делать таблицы, использующие для интерполяции чебышевские многочлены.

Чебышевские приближения и приближения по наименьшим квадратам могут быть объединены в рамках одной идеи. Предположим, что мы минимизируем функцию

$$\min_k = \left(\sum_{i=1}^N |\varepsilon_i|^k \right)^{\frac{1}{k}}.$$

Для $k = 2$ это — метод наименьших квадратов, для $k = \infty$ — метод Чебышева. Этот факт приводит к другой точке зрения на чебышевское приближение. Найдем сначала приближение методом наименьших квадратов. Затем, используя квадрат ошибки как весовую функцию, снова найдем приближение по наименьшим квадратам. Четвертую степень новой ошибки используем как весовую функцию и повторим все снова. Таким способом мы постепенно приблизимся к m_∞ .

Разложение по многочленам Чебышева дает взвешенное приближение по наименьшим квадратам. Если это разложение сходится достаточно быстро, то первый отброшенный член можно рассматривать как ошибку чебышевского приближения. В любом случае такое разложение дает приближение в чебышевском смысле (см. § 19.3).

§ 19.10. Критерий совпадения моментов

До сих пор мы рассматривали три критерия выбора из класса многочленов конкретной функции для использования ее вместо данной в аналитических операциях:

- 1) *создание в узловых точках,*
- 2) *наименьшие квадраты,*
- 3) *чебышевское или минимаксное.*

Еще один критерий, который уже был использован в § 2.9 и который широко применяется в статистике, — это подбор приближающей функции по совпадению нескольких моментов. В статистике употребляются обычно центральные моменты (кроме первого), но если все моменты до k -го совпадают, то и центральные моменты тоже совпадают, так что мы получим один и тот же результат, используются ли «моменты» или «центральные моменты», хотя ошибки округления могут быть совсем разными.

Эта концепция представляет великолепный пример способа выбора критерия. В статистике моменты используются часто; отсюда ясно, почему совпадение моментов при подстановке одной функции вместо другой есть, вероятно, хорошо выбранный критерий для статистических задач.

В качестве примера метода совпадения моментов рассмотрим аппроксимацию функции

$$y(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

квадратным трехчленом. Как правило, $y(x)$ представляет распределение вероятностей и, конечно,

$$\int_{-\infty}^{\infty} y(x) dx = 1.$$

Если взять два первых члена разложения в степенной ряд (только там, где парабола выше оси x), то полная вероятность приближения не была бы равна 1.

Таким образом, мы должны отказаться от приближения степенным рядом и выбрать параболу вида

$$y_0 = A (1 - b^2 x^2) \quad \left(-\frac{1}{b} \leq x \leq \frac{1}{b} \right)$$

по совпадению моментов.

Моменты первоначального распределения

$$\frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{+\infty} x^k e^{-\frac{x^2}{2\sigma^2}} dx = m_k$$

дают

$$m_0 = 1, \quad m_1 = 0, \quad m_2 = \sigma^2,$$

Для приближения получаем

$$\begin{aligned} m_k &= A \int_{-1/b}^{1/b} x^k (1 - b^2 x^2) dx = A \left(\frac{x^{k+1}}{k+1} - b^2 \frac{x^{k+3}}{k+3} \right) \Big|_{-1/b}^{1/b} = \\ &= \frac{2A [1 + (-1)^k]}{b^{k+1} (k+1) (k+3)}. \end{aligned}$$

Отсюда

$$\bar{m}_0 = \frac{4A}{3b}, \quad \bar{m}_1 = 0, \quad \bar{m}_2 = \frac{4A}{15b^3};$$

следовательно,

$$1 = \frac{4A}{3b}, \quad \sigma^2 = \frac{4A}{15b^3}.$$

Отсюда

$$b = \frac{1}{\sigma\sqrt{5}}, \quad A = \frac{3}{4\sigma\sqrt{5}}.$$

Итак, парабола

$$y = \frac{3}{4\sigma\sqrt{5}} \left[1 - \left(\frac{x^2}{5\sigma^2} \right)^2 \right]$$

приближает нормальное распределение, сохраняя нулевой, первый и второй моменты.

ГЛАВА 20

РАЦИОНАЛЬНЫЕ ФУНКЦИИ

§ 20.1. Введение

До сих пор предполагалось, что между узловыми точками функция ведет себя как многочлен. В этой главе мы будем предполагать, что функция ведет себя как отношение двух многочленов, обычно называемое *рациональной функцией*.

Рациональная функция обладает фундаментальным свойством оставаться рациональной при переносе и растяжении независимой переменной. Рациональные функции обладают рядом свойств, часто требуемых от приближающей функции, которой мы собираемся заменять данную в аналитических операциях. Наиболее важно, что рациональными функциями можно приближать такие, которые принимают бесконечные значения для конечных значений аргумента.

Другое свойство — то, что ими можно приближать прямые линии, главным образом ось Ox для больших значений x , чего нельзя сделать нетривиальным многочленом. Рациональные функции легко и

быстро считаются вычислительной машиной. Поэтому их часто используют, чтобы получить приближение для вычисления более сложной функции. Для этих целей, вероятно, нужен чебышевский тип минимаксных приближений.

Теория аппроксимации рациональными функциями находится в запутанном, но быстро развивающемся состоянии; мы дадим о ней только самое общее представление. Одна из причин, по которой мы не будем слишком глубоко заходить в нее, — та, что среднему вычислителю едва ли придется сделать в год много рациональных приближений, и более эффективный метод, позволяющий сберечь несколько секунд работы машины, по-видимому, не стоит в элементарных случаях затраты времени вычислителя.

§ 20.2. Непосредственный подход

Пусть дана функция, по которой можно вычислять узлы (x_i, y_i) или сами узлы и требуется подобрать по этим данным рациональную функцию,

$$y = f(x) \approx \frac{N(x)}{D(x)},$$

где $N(x)$ и $D(x)$ — многочлены. Сначала решим, какого вида многочлены следует искать. Пусть в частном случае нанесение на график точек $(\log x_i, \log y_i)$ дало для малых $x_i > 0$ наклон, примерно равный 3, откуда мы заключили, что надо попробовать

$$y = f(x) \approx \frac{ax^3 + \dots}{1 + \dots}.$$

Для больших x точки (x_i, y_i) легли примерно на прямую с положительным наклоном. Это значит, что степень $N(x)$ должна быть на 1 больше степени $D(x)$. Обозначим степень $D(x)$ через k . Тогда

$$y = f(x) = \frac{a_3 x^3 + a_4 x^4 + \dots + a_{k+1} x^{k+1}}{1 + b_1 x + b_2 x^2 + \dots + b_k x^k}. \quad (20.2.1)$$

Остается выбрать показатель степени k . Трудно сказать, как именно выбирать k ; единственное, что можно посоветовать — принять во внимание получающееся число параметров относительно «сложности» (субъективно) данных.

Когда k выбрано, становится известным число параметров. Заметим, что можно фиксировать один коэффициент, так как встречается только отношение $N(x)$ и $D(x)$, следовательно, один коэффициент (заведомо ненулевой) должен быть зафиксирован*). Мы выбираем

*) Практически обычно лучше выбрать единицей старший коэффициент числителя или знаменателя: это экономит одно умножение при вычислении рациональной функции.

или вычисляем соответствующее число M узлов, расположенных там, где это кажется наиболее важным. Тогда имеем

$$D(x_i)y_i = N(x_i) \quad (i = 1, \dots, M),$$

систему M уравнений, которую надо решить. В примере (равенство (20.2-1)) имеем

$$y_i + b_1 x_i y_i + b_2 x_i^2 y_i + \dots + b_k x_i^k y_i = a_3 x_i^3 + a_4 x_i^4 + \dots + a_{k+1} x_i^{k+1} \\ (i = 1, 2, \dots, 2k - 1). \quad (20.2-2)$$

Это — линейные уравнения относительно неизвестных a_i и b_i , которые могут быть решены при помощи обычных библиотечных программ.

Часто встречается мнение, что эта система высокого порядка. Насколько можно судить, число неизвестных редко превышает 10—12, так что решение требует не более 1—2 тысяч операций; такая работа едва ли сложна для большой машины, если это требуется сделать один раз на задачу.

Прежде чем принять аппроксимирующую функцию, полезно вычислить достаточно много ее значений. Однажды, в практике автора, знаменатель имел корень между двумя узлами, тем самым обращая в бесконечность приближающую функцию. Между теми же двумя узлами у числителя тоже был корень, чем маскировалось поведение функции.

Если приближение неудовлетворительно, то все, что можно попробовать, сводится к следующему:

- 1) сдвинуть узловые точки;
- 2) изменить вид приближающей функции;
- 3) изменить k .

Выбор одного из способов действий зависит от того, почему именно была неудовлетворительна аппроксимация.

Упражнения

20.2-1. Какой вид рациональной функции следует выбрать, если $y(x) \geq 0$ и $y(x) \rightarrow 0$ при $x \rightarrow \infty$?

20.2-2. Какой вид рациональной функции выбрать, если $y(x)$ имеет корень при $x = a$; нуль $D(x)$ при $x = b$; и то и другое?

§ 20.3. Чебышевское приближение рациональными функциями

Как указано в § 20.1, рациональная функция часто используется как легко вычисляемое приближение трансцендентной функции. Когда такое приближение должно использоваться много раз, или в часто повторяющейся задаче, или в специализированной машине, то скорее всего желательно чебышевское равноколеблющееся (минимаксное) приближение.

Начнем с выбора набора узлов x_i и вычислим соответствующие y_i . По ним находим рациональную функцию подходящего вида (как в § 20.2).

Затем рисуем кривую ошибки, вычисляя ошибки во многих точках; там, где ошибка велика, сдвинем точки поближе друг к другу, там же, где локальная максимальная ошибка мала, раздвинем точки подальше. С этими вновь выбранными узлами повторим весь процесс. Число повторений, необходимых для получения почти равноколеблющейся кривой ошибки, редко доходит до десяти. Скорость приближения к равноколеблющейся кривой ошибки зависит, среди прочего, от способа перемещения узлов.

Мы не предлагаем никаких доказательств того, что этот метод будет работать, но практика показывает, что ряд простых схем работает прилично. Чтобы не потерять слишком много знаков, обычно требуется несколько «маленьких хитростей» при организации вычислений.

§ 20.4. Обратные разности (симметричные)

В предыдущем параграфе использовалась итеративная схема, требующая нескольких повторений, каждое из которых содержит решение системы линейных уравнений (20.2-2). Это заставляет пытаться найти регулярный способ решения такой системы.

Следуя Милну*), рассмотрим частный случай

$$y = \frac{a_0 + a_1x + a_2x^2 + a_3x^3}{b_0 + b_1x + b_2x^2},$$

который использует шесть заданных точек (x_i, y_i) ($i = 1, \dots, 6$).

Приводим к общему знаменателю:

$$a_0 - b_0y + a_1x - b_1xy + a_2x^2 - b_2x^2y + a_3x^3 = 0.$$

Применяя метод, который был использован для вывода формул (8.2-3), легко увидеть, что определитель

$$\begin{vmatrix} 1 & y & x & xy & x^2 & x^2y & x^3 \\ 1 & y_1 & x_1 & x_1y_1 & x_1^2 & x_1^2y_1 & x_1^3 \\ 1 & y_2 & x_2 & x_2y_2 & x_2^2 & x_2^2y_2 & x_2^3 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & y_6 & x_6 & x_6y_6 & x_6^2 & x_6^2y_6 & x_6^3 \end{vmatrix} = 0 \quad (20.4-1)$$

дает требуемое решение.

Приведем сначала вторую строку к виду $(1, 0, 0, 0, 0, 0, 0)$. Умножаем:

1. Первый столбец на y_1 и вычитаем его из второго столбца;
2. Третий столбец на y_1 и вычитаем его из четвертого столбца;
3. Пятый столбец на y_1 и вычитаем его из шестого столбца;

*) См. [25]. Мы немного изменили обозначения.

4. Пятый столбец на x_1 и вычитаем его из седьмого столбца;
5. Третий столбец на x_1 и вычитаем его из пятого столбца;
6. Первый столбец на x_1 и вычитаем его из третьего столбца;

Разложим затем получившийся определитель по второй строке; и разделим

7. Первую строку на $y - y_1$;
8. Вторую строку на $y_2 - y_1$;
- третью строку на $y_3 - y_1$;
-
9. Седьмую строку на $y_6 - y_1$.

При делении могут возникнуть неприятности, если какой-нибудь $y_i = y_1$ ($i \neq 1$) и даже если они близки. В действительности строку, которая будет приводиться к виду $(1, 0, 0, 0, 0, 0, 0)$, следует выбирать, имея в виду это деление и стараясь избежать большой интерференции в разности $y_i - y_1$ (и тем самым деления на маленькое, неточное число). Обозначим

$$\frac{x - x_1}{y - y_1} = \rho_1(x, x_1), \quad \frac{x_i - x_1}{y_i - y_1} = \rho_1(x_i, x_1).$$

Определитель (20.4-1) запишется теперь так:

$$\begin{vmatrix} 1 & \rho_1(x, x_1) & x & x\rho_1(x, x_1) & x^2 & x^2\rho_1(x, x_1) \\ 1 & \rho_1(x_2, x_1) & x_2 & x_2\rho_1(x_2, x_1) & x_2^2 & x_2^2\rho_1(x_2, x_1) \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & \rho_1(x_6, x_1) & x_6 & x_6\rho_1(x_6, x_1) & x_6^2 & x_6^2\rho_1(x_6, x_1) \end{vmatrix} = 0, \quad (20.4-2)$$

т. е. останется того же вида, что и раньше, но только шестого порядка.

Повторим процедуру. Умножаем:

1. Первый столбец на $\rho_1(x_2, x_1)$ и вычитаем его из второго столбца;
2. Третий столбец на $\rho_1(x_2, x_1)$ и вычитаем его из четвертого столбца;
3. Пятый столбец на $\rho_1(x_2, x_1)$ и вычитаем его из шестого столбца;

4. Третий столбец на x_2 и вычитаем его из пятого столбца;
5. Первый столбец на x_2 и вычитаем его из третьего столбца;

Разлагаем определитель и делим

первую строку на $\rho_1(x, x_1) - \rho_1(x_2, x_1)$,

вторую строку на $\rho_1(x_3, x_1) - \rho_1(x_2, x_1)$.

Здесь разумный выбор строки, приводимой к виду $(1, 0, 0, 0, 0, 0)$, может помочь сохранить точность.

Если мы обозначим

$$u_x = \frac{x - x_2}{\rho_1(x, x_1) - \rho_1(x_2, x_1)},$$

то получим для (20.4-2) выражение

$$\begin{vmatrix} 1 & u_x & x & x & u_x & x^2 \\ 1 & u_3 & x_3 & x_3 & u_3 & x_3^2 \\ 1 & u_4 & x_4 & x_4 & u_4 & x_4^2 \\ 1 & u_5 & x_5 & x_5 & u_5 & x_5^2 \\ 1 & u_6 & x_6 & x_6 & u_6 & x_6^2 \end{vmatrix} = 0. \quad (20.4-3)$$

Функция $\rho_1(x_i, x_j)$ была, очевидно, симметрична относительно своих переменных. Но это неверно для функции, которую мы обозначили u_i . Обычно используют симметричные функции

$$\begin{aligned} & \frac{x - x_2}{(x - x_1)/(y - y_1) - (x_2 - x_1)/(y_2 - y_1)} + y_1 = \\ &= \frac{x - x_1}{(x - x_1)/(y - y_1) - (x_2 - x_1)/(y_2 - y_1)} + y_1 = \\ &= \frac{x_2 - x_1}{(x_2 - x)/(y_2 - y) - (x - x_1)/(y - y_1)} + y \end{aligned}$$

и обозначают их $\rho_2(x, x_1, x_2)$.

Чтобы привести определитель к нужному виду, умножим первый столбец на y_1 и прибавим его ко второму столбцу, третий столбец — на y_1 и прибавим его к четвертому столбцу. Имеем

$$\begin{vmatrix} 1 & \rho_2(x, x_2, x_1) & x & x\rho_2(x, x_2, x_1) & x^2 \\ 1 & \rho_2(x_3, x_2, x_1) & x_3 & x_3\rho_2(x_3, x_2, x_1) & x_3^2 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & \rho_2(x_6, x_2, x_1) & x_6 & x_6\rho_2(x_6, x_2, x_1) & x_6^2 \end{vmatrix} = 0. \quad (20.4-4)$$

Вопрос ценности этой симметрии для практических вычислений остается открытым, но теория от этого выглядит проще и мы будем придерживаться симметричной формы.

Повторяя процесс, получим, наконец,

$$\begin{vmatrix} 1 & \rho_5(x, x_5, x_4, x_3, x_2, x_1) \\ 1 & \rho_5(x_6, x_5, x_4, x_3, x_2, x_1) \end{vmatrix} = 0. \quad (20.4-5)$$

Точно так же, как было установлено, что формула Ньютона есть тождество, можно написать

$$\begin{aligned} y &= y_1 + \frac{x - x_1}{\rho_1(x, x_1)} = y_1 + \frac{x - x_1}{\rho_1(x_2, x_1) + (x - x_2)/[\rho_2(x, x_2, x_1) - y_1]} = \\ &= y_1 + (x - x_1) \{ \rho_1(x_2, x_1) + (x - x_2)/[\rho_2(x_2, x_2, x_1) - y_1] + \\ &+ (x - x_3)/[\rho_3(x_1, x_3, x_2, x_1) - \rho_1(x_2, x_1)] \}^{-1} = \dots \quad (20.4-6) \end{aligned}$$

В данном частном случае мы заканчиваем членом

$$\frac{x - x_5}{\rho_5(x_6, x_5, x_4, x_3, x_2, x_1) - \rho_3(x_4, x_3, x_2, x_1)},$$

так как

$$\rho_5(x, x_5, x_4, x_3, x_2, x_1) = \rho_3(x_6, x_5, x_4, x_3, x_2, x_1).$$

Величины ρ_i называются обратными разностями; удобно представлять себе таблицу обратных разностей в виде

$$\begin{array}{lcl} x_1 & y_1 & \\ & \rho_1(x_2, x_1) & \\ x_2 & y_2 & \rho_2(x_3, x_2, x_1) \\ & \rho_1(x_3, x_1) & \rho_3(x_4, x_3, x_2, x_1) \\ x_3 & y_3 & \rho_2(x_4, x_2, x_1) \\ & \rho_1(x_4, x_1) & \\ x_4 & y_4 & \\ . & . & . \end{array}$$

§ 20.5. Пример

Рассмотрим часто встречающуюся функцию

$$y = \frac{1}{1+x^2}$$

и построим таблицу ее значений. Как и для разделенных разностей, вместо того чтобы использовать первую и k -ю строки, можно брать разности смежных строк; пользоваться будем только значениями верхней диагонали. Итак, получаем таблицу 20.5-1.

Т а б л и ц а 20.5-1

x	y				
0	1				
1	1/2	-2			
2	1/5	-10/3	-1	0	
3	1/10	-50/5	-1/10	40	0
4	1/17	-170/7	-1/25	140	0
5	1/26	-442/9	-1/46	324	
6	1/37	-962/11	-1/73		

Столбец нулей (или почти нулей) играет ту же роль, что и в обычной таблице разностей; он показывает, где надо остановиться. Теперь из (20.4-6) прямой подстановкой получаем

$$y = 1 + \frac{x-0}{-2 + \frac{x-1}{-2 + \frac{x-2}{2 + \frac{x-3}{1}}}},$$

что при переходе от непрерывной дроби к рациональной функции дает

$$y = \frac{1}{1+x^2}.$$

Мы не будем углубляться в эту область и просто отошлем читателя к образцовым руководствам Милна [25], Милн-Томсона [27] и Хильдебранда [14].

НЕМНОГОЧЛЕННЫЕ ПРИБЛИЖЕНИЯ

ГЛАВА 21

ПЕРИОДИЧЕСКИЕ ФУНКЦИИ. АППРОКСИМАЦИЯ ФУРЬЕ

§ 21.1. Цель этой теории

Процесс решения любой практической задачи на машине включает три этапа: планирование, выполнение плана и интерпретацию результатов. Эти три этапа можно сравнить с тремя классическими фазами шахматной игры: дебют, миттельшпиль и эндшпиль. В шахматах все три фазы имеют ясно выраженный характер, хотя каждая из них незаметно переходит в следующую и определяется предыдущей, и может случиться, что игра окончится в миттельшпиле или даже в дебюте. Так же дело обстоит и в вычислительной практике.

На первом этапе все расчеты носят, как правило, характер набросков, в которых определяется объем вычислений, машинное время и т. д. Обычно на этапе планирования принимаются решения, подобные следующему: «Сначала вычислим функцию в некоторой группе точек, затем заменим интеграл в интегральном уравнении гауссовой квадратурой по семи точкам, решим полученную систему уравнений методом исключения, подставим полученные решения в такие-то и такие-то формулы и напечатаем нужные результаты». Начальный этап включает в себя оценки времени программирования, кодировки, машинного времени, оценки того, когда будут получены результаты и как они будут использованы.

Обычно в процессе решения задачи возникает обратная связь между постановкой задачи и вычислением, и поэтому точная постановка всей задачи возможна лишь после начала вычислений. Тем не менее выходить на машину нужно, лишь тщательно выполнив этап планирования.

На этапе «выполнения» план часто меняется. Например, после вычисления упомянутой функции становится ясным, что вместо семиточечной гауссовой формулы следует использовать десятиточечную, а это может вдвое увеличить количество работы для решения получающейся системы уравнений. Ошибки при решении уравнений могут оказаться большими, и тогда для улучшения полученных решений может потребоваться проведение итераций, что усложнит программу и увеличит машинное время.

В фазе «интерпретации» необходимо не только обсудить и объяснить результаты, но также и проверить, соответствуют ли результаты физической модели и не является ли часть из них следствием формально проведенных вычислений, а не физических закономерностей. Кроме того, нужно объяснить все изменения в плане вычислений. Почему оказалась сильно отличающейся от предполагаемой функция фактически «увиденная»? Почему уравнения оказались трудны для решения? Значит ли это, что результаты очень чувствительны к исходной функции или к некоторым параметрам? Все эти вопросы требуют исследования.

Обычно в математических кругах пренебрегают первой и третьей фазами, как не относящимися к математике. Но эти две фазы, в особенности третья, которой пренебрегают чаще всего, имеют решающее значение для успеха всей работы.

Полиномиальные методы, которые рассматривались во второй части книги, не приспособлены для исследований в фазах планирования и интерпретации. Ошибки в них выражаются через производные высоких порядков и, следовательно, на этапе планирования редко могут быть оценены точно; более поздняя оценка их на этапе интерпретации тоже не всегда проливает свет на первоначальную проблему.

Целью этой и следующих четырех глав является изложение элементов теории аппроксимации функциями с ограниченным спектром, которая для широкого класса задач дает модели, позволяющие математику действовать в какой-то мере автоматически на этапах планирования и интерпретации. В результате в фазе планирования можно сделать реалистическую оценку расходов и потребного времени, а то, что произойдет в процессе вычислений, может пролить новый свет на изучаемую физическую ситуацию. Теория изложена не полно; она отвечает не на все вопросы и иногда требует несколько больше вычислений, чем многочленная модель. Однако опыт работы с ней в течение нескольких лет показывает, что часто модель функции с ограниченным спектром оказывает большую помощь как во время планирования вычислительной работы, так и во время интерпретации результатов. Повторяем девиз этой книги:

Цель расчетов — понимание, а не числа.

§ 21.2. Замена переменных и выбор узлов

В большинстве вычислений узловые значения ищутся на множестве равноудаленных точек. Пусть расстояние между ними принято за единицу, т. е. $h = \Delta t = 1$. Исследуем сначала эффективность такого выбора для синусоидальных функций.

Для синусоидальной функции вида $\cos [\pi (n + \varepsilon) t + \varphi]$ существует другая синусоидальная функция $\cos [\pi (n - \varepsilon) t - \varphi]$, которая имеет одинаковые с ней узловые значения в точках $t = 0, 1, 2, \dots$. Это

видно из тригонометрического тождества

$$\begin{aligned}\cos [\pi (n + \varepsilon) t + \varphi] - \cos [\pi (n - \varepsilon) t - \varphi] = \\ = -2 \sin (\pi n t) \sin (\pi \varepsilon t + \varphi) = 0\end{aligned}$$

(для целых t и n). Таким образом, при вычислении в этих точках частоты $\pi (n + \varepsilon)$ и $\pi (n - \varepsilon)$ ведут себя так, как будто они одинаковы при условии, что фазовые углы надлежащим образом связаны друг с другом.

Этот эффект хорошо известен кино- и телезрителям. В обоих случаях движущаяся картина фиксируется 20 раз в секунду. Когда колесо вагона начинает вращаться, зритель сперва видит, что оно крутится все быстрее и быстрее; но затем движение начинает замедляться, колесо останавливается и, наконец, начинает вращаться в обратную сторону. При дальнейшем возрастании скорости кажется, будто колесо опять начинает вращаться вперед, замедляется, останавливается, крутится в обратную сторону и т. д.

Этот хорошо известный стробоскопический эффект часто используется для анализа периодических явлений. Устраивают так, чтобы вспышки света появлялись с частотой немного меньшей, чем период этого явления. Таким образом, каждая вспышка показывает периодическое явление в чуть более поздней фазе его цикла. Существует много частот, которые могут быть использованы для стробоскопического освещения, дающего одинаковые результаты для зрителя. Так, если при одной частоте наблюдается определенная кажущаяся скорость вращения, то при частоте, равной приблизительно половине этой, человеческий глаз будет наблюдать ту же самую скорость, однако качество картины несколько ухудшается из-за низкочастотного мерцания.

В этих двух примерах человеческий глаз видит две компоненты движения x и y , следовательно, может отличить прямое движение от обратного. Если он видит только одну компоненту, то в некоторых случаях движение вперед неотличимо от движения назад. Здесь происходит смешение частот, когда несколько различных частот дают один и тот же результат во всех рассматриваемых точках. Очевидно, что это — неизбежное следствие выборки с равными интервалами; этим фактом не следует пренебрегать.

Упражнение 21.2-1. Нарисуйте графики функций $\cos(1,1\pi t)$ и $\cos(0,9\pi t)$ для $0 \leq t \leq 5$ и рассмотрите поведение их в узловых точках $t = 0, 1, 2, 3, 4, 5$.

§ 21.3. Ряды Фурье; периодические явления

Если $y = y(t)$ — периодическая функция от t с периодом $2N$, то теория рядов Фурье утверждает (см. гл. 6), что

$$y(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} \left(a_k \cos \frac{\pi}{N} kt + b_k \sin \frac{\pi}{N} kt \right) \quad (21.3-1)$$

при условии, что на $y(t)$ наложены довольно слабые (с точки зрения вычислительной математики) ограничения. Величины a_k и b_k определяются по формулам:

$$\left. \begin{aligned} a_k &= \frac{1}{N} \int_0^{2N} y(t) \cos \frac{\pi}{N} kt dt, \\ b_k &= \frac{1}{N} \int_0^{2N} y(t) \sin \frac{\pi}{N} kt dt. \end{aligned} \right\} \quad (21.3-2)$$

В гл. 6 были приведены конечные ряды Фурье, построенные по значениям в $2N$ точках. Если вычислить конечный ряд Фурье, построенный по $2N$ узловым точкам $0, 1, 2, \dots, 2N-1$, то получим

$$y(t) \sim \frac{A_0}{2} + \sum_{k=1}^{N-1} A_k \cos \frac{\pi}{N} kt + \sum_{k=1}^{N-1} B_k \sin \frac{\pi}{N} kt + \frac{A_N}{2} \cos \pi t, \quad (21.3-3)$$

где

$$A_k = \frac{1}{N} \sum_{t=0}^{2N-1} y(t) \cos \frac{\pi}{N} kt, \quad B_k = \frac{1}{N} \sum_{t=0}^{2N-1} y(t) \sin \frac{\pi}{N} kt. \quad (21.3-4)$$

Покажем теперь, что между дискретными и непрерывными коэффициентами существуют следующие соотношения:

$$\left. \begin{aligned} A_0 &= a_0 + 2 \sum_{j=1}^{\infty} a_{2Nj} \\ A_k &= a_k + \sum_{j=1}^{\infty} (a_{2Nj-k} + a_{2Nj+k}), \\ B_k &= b_k + \sum_{j=1}^{\infty} (-b_{2Nj-k} + b_{2Nj+k}). \end{aligned} \right\} \quad (21.3-5)$$

Для доказательства просуммируем обе части (21.3-1) при $t=0, 1, 2, \dots, 2N-1$ и получим

$$\sum_{t=0}^{2N-1} y(t) = \frac{a_0}{2} \sum_{t=0}^{2N-1} 1 + \sum_{k=1}^{\infty} a_k \sum_{t=0}^{2N-1} \cos \frac{\pi}{N} kt + \sum_{k=1}^{\infty} b_k \sum_{t=0}^{2N-1} \sin \frac{\pi}{N} kt.$$

Если $k \neq 0, 2N, 4N, \dots$, то

$$\sum_{t=0}^{2N-1} \cos \frac{\pi}{N} kt = 0, \quad \sum_{t=0}^{2N-1} \sin \frac{\pi}{N} kt = 0.$$

При $k=0, 2N, 4N, \dots$, полагая $k=2N_j$, получим

$$\sum_{t=0}^{2N-1} \cos \frac{\pi}{N} (2N_j) t = \sum_{t=0}^{2N-1} 1 = 2N, \quad \sum_{t=0}^{2N-1} \sin \frac{\pi}{N} (2N_j) t = \sum_{t=0}^{2N-1} 0 = 0.$$

Отсюда, используя (21.3-2) и (21.3-4), имеем

$$2N \frac{A_0}{2} = 2N \frac{a_0}{2} + 2N \sum_{j=1}^{\infty} a_{2Nj},$$

что совпадает с первым из уравнений (21.3-5).

Если умножить (21.3-1) на $\cos \frac{\pi m}{N} t$ ($0 < m < N$) и просуммировать, то аналогично получим

$$\begin{aligned} NA_m &= \sum_{k=1}^{\infty} a_k \sum_{t=0}^{2N-1} \cos \frac{\pi k}{N} t \cos \frac{\pi m}{N} t = \\ &= \sum_{k=1}^{\infty} a_k \sum_{t=0}^{2N-1} \frac{1}{2} \left[\cos \frac{\pi(k+m)t}{N} + \cos \frac{\pi(k-m)t}{N} \right], \end{aligned}$$

откуда следует второе из уравнений (21.3-5). Третье уравнение получается умножением на $\sin \frac{\pi m}{N} t$ и суммированием.

Соотношение (21.3-5) между коэффициентами разложений на непрерывном и дискретном множествах точек ясно показывает явление,

которое можно назвать «мимикрия частот». Если мы представим частоты в виде точек бесконечной прямой, то явление «мимикрии» заключается в том, что прямая как бы складывается гармошкой. Первая частота, на которой происходит перегиб прямой, называется

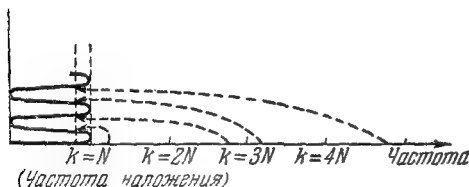


Рис. 21.3-1. Смещение частот.

частотой наложения или частотой Найквиста и соответствует значению $k=N$. Последующие наложения будут происходить через такой же промежуток. Все точки кривой (рис. 21.3-1), расположенные над одной и той же точкой оси частот, в результате выборки проявляют себя как одна частота. Однажды произведя выборку, мы уже не сможем снова разделить частоты, которые «мимикрировали»

друг под друга. На самую частоту наложения мы попадаем один раз в два цикла, когда $\sin \pi t = 0$, $\cos \pi t = (-1)^t$ (t — целое).

Упражнения

21.3-1. Получите третье уравнение из 21.3-5.

21.3-2. Найдите уравнение для A_n из 21.3-5.

§ 21.4. Интерполяция периодических функций

Во второй части была изложена теория многочленной аппроксимации. Займемся теперь изложением подобной теории для аппроксимации периодических функций рядами Фурье.

Несмотря на то, что задача может быть поставлена для множества неравноудаленных точек и известны некоторые результаты для этого случая, однако и теория и практика обычно имеют дело с множеством равноудаленных точек. Для случая $2N$ равноудаленных узловых точек задача интерполяции решается путем применения конечного ряда Фурье (гл. 6). Значение функции в промежутках между узлами вычисляется с помощью ряда.

Как и во второй части, после получения формулы необходимо определить ошибку аппроксимации. Эту ошибку можно разделить на две части: ошибку, возникающую вследствие использования конечного числа членов ряда, и ошибку вследствие дискретного задания функции.

Влияние дискретности было уже исследовано; остается влияние конечности числа членов ряда Фурье.

Если непрерывный ряд Фурье имеет вид (21.3-1), то, подставляя формулы коэффициентов (21.3-2), получим первые $2N$ членов, из которых, как и в случае дискретного ряда, возьмем половину косинусного члена

$$\begin{aligned} y_{2N}(t) &= \frac{1}{N} \int_0^2 y(s) \left\{ \frac{1}{2} + \sum_{k=1}^{N-1} \left[\cos \frac{\pi}{N} kt \cos \frac{\pi}{N} ks + \sin \frac{\pi}{N} kt \sin \frac{\pi}{N} ks \right] + \right. \\ &\quad \left. + \frac{1}{2} \cos \pi t \cos \pi s \right\} ds = \\ &= \frac{1}{N} \int_0^{2N} y(s) \left\{ \frac{1}{2} + \sum_{k=1}^{N-1} \left[\cos \frac{\pi}{N} k(t-s) \right] + \frac{1}{2} \cos \pi(t-s) \right\} ds. \end{aligned}$$

Чтобы просуммировать величину*), стоящую в фигурных скобках,

*) Этот ряд можно просуммировать более элегантно, используя методы гл. 3. Однако читатель, вероятно, успел забыть детали, и поэтому мы будем пользоваться более грубыми, но и более знакомыми тригонометрическими методами.

умножим ее на $\sin \frac{\pi}{2N}(t-s)$; в результате получим

$$\begin{aligned} \frac{1}{2} \left\{ \sin \frac{\pi}{2N}(t-s) + \left[\sin \frac{\pi}{N} \left(1 + \frac{1}{2} \right) (t-s) - \sin \frac{\pi}{N} \left(1 - \frac{1}{2} \right) (t-s) \right] + \right. \\ \left. + \left[\sin \frac{\pi}{N} \left(2 + \frac{1}{2} \right) (t-s) - \sin \frac{\pi}{N} \left(2 - \frac{1}{2} \right) (t-s) \right] + \right. \\ \dots \dots \dots \left. + \left[\sin \frac{\pi}{N} \left(N-1 + \frac{1}{2} \right) (t-s) - \sin \frac{\pi}{N} \left(N-1 - \frac{1}{2} \right) (t-s) \right] + \right. \\ \left. + \sin \frac{\pi}{2N}(t-s) \cos \pi(t-s) \right\}. \quad (21.4-1) \end{aligned}$$

Все члены, кроме двух, уничтожаются:

$$\frac{1}{2} \left\{ \sin \left[\frac{\pi}{N} \left(N - \frac{1}{2} \right) (t-s) \right] + \sin \frac{\pi}{2N}(t-s) \cos \pi(t-s) \right\}.$$

Распишем первый из этих членов по формуле

$$\begin{aligned} \sin \left[\frac{\pi}{N} \left(N - \frac{1}{2} \right) (t-s) \right] = \sin \pi(t-s) \cos \frac{\pi}{2N}(t-s) - \\ - \cos \pi(t-s) \sin \frac{\pi}{2N}(t-s). \end{aligned}$$

В результате второй член последней формулы уничтожится со вторым членом предпоследней, и выражение в скобках примет вид

$$\frac{\sin \pi(t-s) \cos \frac{\pi}{2N}(t-s)}{2 \sin \frac{\pi}{2N}(t-s)},$$

тогда

$$y_{2N}(t) = \frac{1}{N} \int_0^{2N} y(s) \frac{\sin \pi(t-s) \cos \frac{\pi}{2N}(t-s)}{2 \sin \frac{\pi}{2N}(t-s)} ds. \quad (21.4-2)$$

Поскольку $y(s)$ — периодическая функция, то отрезок интегрирования можно сдвинуть в любой интервал длины $2N$; кроме того, положим $t-s = \theta$:

$$y_{2N}(t) = \frac{1}{2N} \int_{-N}^N y(t-\theta) \frac{\sin \pi \theta \cos \frac{\pi}{2N} \theta}{\sin \frac{\pi}{2N} \theta} d\theta. \quad (21.4-3)$$

Если $y(t) = 1$, то

$$1 = \frac{1}{2N} \int_{-N}^N \frac{\sin \pi \theta \cos \frac{\pi \theta}{2N}}{\sin \frac{\pi \theta}{2N}} d\theta.$$

Но, поскольку интегрирование ведется по θ , то

$$y(t) = \frac{1}{2N} \int_{-N}^N y(t) \frac{\sin \pi \theta \cos \frac{\pi \theta}{2N}}{\sin \frac{\pi \theta}{2N}} d\theta. \quad (21.4-4)$$

Вычитая (21.4-3) из (21.4-4), получим ошибку от использования $2N$ членов ряда

$$\epsilon_{2N} = y(t) - y_{2N}(t) = \frac{1}{2N} \int_{-N}^N [y(t) - y(t - \theta)] \frac{\sin \pi \theta \cos \frac{\pi \theta}{2N}}{\sin \frac{\pi \theta}{2N}} d\theta. \quad (21.4-5)$$

Любопытно отметить, что в большинстве работ по рядам Фурье суммируется нечетное число членов, и поэтому вместо (21.4-5) получается похожая, но менее удобная формула *). Но простота формулы (21.4-5) обманлива: чем больше узлов взять, тем длиннее будет отрезок интегрирования. Однако на практике отрезок имеет фиксированную длину и изменяет интервал между узловыми точками. Поэтому мы производим замену переменных и обозначаем (что может привести к некоторой путанице):

$$\begin{aligned} \theta &= 2N\varphi, & t &= 2N\gamma, & y(2Nx) &= y(x), \\ \epsilon_{2N} &= \int_{-1/2}^{1/2} [y(\gamma) - y(\gamma - \varphi)] \left\{ \frac{\sin 2N\pi\varphi \cos \pi\varphi}{\sin \pi\varphi} \right\} d\varphi. \end{aligned} \quad (21.4-6)$$

Величину в фигурных скобках часто называют ядром $K(\varphi)$. Очевидно, что $K(0) = 2N$. При больших N член $\sin(2N\pi\varphi)$ меняет знак, а $\operatorname{ctg} \pi\varphi$ определяет затухание этих колебаний так, что $K(\pm 1/2) = 0$. Огибающая колебаний есть $\operatorname{ctg} \pi\varphi$ и

$$K(\varphi) = K(-\varphi).$$

*) Зигмунд [44] называет полученную формулу «модифицированным ядром».

Если функция $y(x)$ гладкая, то наибольший вклад в интеграл дают значения ядра вблизи $\varphi = 0$, например при $-\frac{1}{2N} \leq \varphi \leq \frac{1}{2N}$

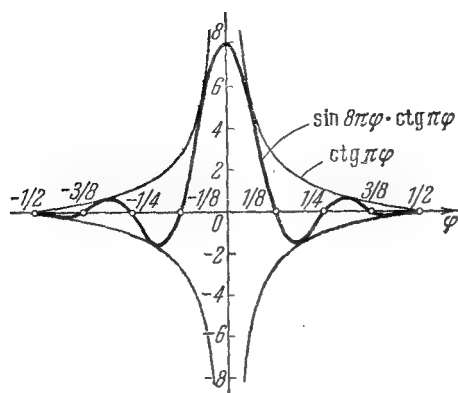


Рис. 21.4-1. Ядро ряда Фурье для $N=4$.

Упражнения

21.4-1. Мы уже аппроксимировали тригонометрические функции полиномами. Приблизьте $y(t) = (1 - t^2)^2$, $(-1 \leq t \leq 1)$, рядом Фурье при $N=3$, т. е. в узловых точках $t = (-\frac{2}{3}; -\frac{1}{3}; 0; \frac{1}{3}; \frac{2}{3}; 1)$. Найдите функцию в точке $t = \frac{1}{2}$.

21.4-2. В этом параграфе выведен эквивалент интерполяции Лагранжа. Исследуйте эквивалент интерполяции Эрмита, которая приближает функцию и производную в каждой узловой точке.

Ответ: Такого не существует.

§ 21.5. Интегрирование

Как и во второй части, исследовав вопрос об интерполировании и его ошибках, перейдем к задаче интегрирования. Если коэффициенты непрерывного ряда Фурье $y(t)$ уменьшаются при $N \rightarrow \infty$ довольно быстро, то коэффициент A_0 конечного ряда Фурье будет давать хорошую оценку интеграла $\int_0^{2N} y(t) dt = Na_0$, поскольку из уравнений (21.3-5) следует

$$A_0 = \frac{1}{N} \sum_{t=0}^{2N-1} y(t) = a_0 + 2 \sum_{j=1}^{\infty} a_{2Nj}.$$

Таким образом, ошибка, возникающая при оценке интеграла на основе равноотстоящих и имеющих равные веса узлов,

дается формулой

$$2N \sum_{j=1}^{\infty} a_j N_j$$

в которой отражается неизбежное смешивание частот — результат дискретности узлов.

Для периодических функций формула интегрирования, использующая $2N$ равноотстоящих и равновзвешенных узлов, соответствует формулам:

Ньютона — Котеса, потому что точки равноотстоящие;

Чебышева, потому что точки имеют равные веса;

Гаусса, потому что $2N$ узлов точно интегрируют $4N$ функций.

Проиллюстрируем точность этой формулы при интегрировании периодических функций. Рассмотрим эллиптические интегралы

$$I_1 = \int_0^{\pi/2} \sqrt{1 - k^2 \sin^2 \theta} d\theta, \quad I_2 = \int_0^{\pi/2} \frac{1}{\sqrt{1 - k^2 \sin^2 \theta}} d\theta,$$

где $|k| \leq 1$ — параметр. Квадратный корень можно представить в виде бинома. Если k^2 мало, то ряд сходится быстро. Выразив $\sin^{2n} \theta$ через $\sin 2p\theta$ и $\cos 2p\theta$ (где $p=0, 1, 2, \dots$), получим быстро сходящийся ряд Фурье

$$\begin{aligned} \left(\frac{e^{i\theta} - e^{-i\theta}}{2i} \right)^{2n} &= \frac{1}{2^{2n} i^{2n}} [e^{2ni\theta} - C(2n, 1)e^{2(n-1)i\theta} + \dots] = \\ &= \frac{1}{2^{2n-1} i^{2n}} [\cos 2n\theta - C(2n, 1) \cos (2n-2)\theta + \dots]. \end{aligned}$$

При малых значениях k^2 результаты будут точными, но при k^2 , близком к 1, нельзя ожидать высокой точности.

Для эксперимента выберем узлы в точках $0^\circ, 30^\circ, 60^\circ, 90^\circ$, хотя для получения периодичности следовало бы выбрать шесть точек на

Таблица 21.5-1
Эллиптический интеграл

k^2	I_1 (выч.)	I_1 (табл.)	Ошибка
$\frac{1}{4}$	1,46746	1,4675	
$\frac{1}{2}$	1,35064	1,3506	
$\frac{3}{4}$	1,21099	1,2111	0,0001

Таблица 21.5-2

k^2	I_2 (выч.)	I_2 (табл.)	Ошибка
$\frac{1}{4}$	1,68575	1,6858	
$\frac{1}{2}$	1,85410	1,8541	
$\frac{3}{4}$	2,15789	2,1565	0,0014

период $0 \leq t \leq \pi$. В действительности мы вычислим значения лишь в четырех узлах, один из которых $\theta = 0$ тривиален. Результаты вычислений и табличные значения этих интегралов приведены в таблицах 21.5-1, 21.5-2.

Из рассмотрения этого примера можно вывести следующее: во-первых, коэффициенты ряда Фурье часто можно оценить с довольно высокой точностью, поэтому можно оценить ошибку вычислений, не приступая к ним;

во-вторых, метод интегрирования, использующий равноотстоящие узлы с равными весами, дает очень точные результаты для периодических функций. Этот метод можно использовать для вычисления функции Бесселя $J_0(x)$ по формуле

$$J_0(x) = \frac{1}{\pi} \int_0^{\pi} \cos(x \sin \varphi) d\varphi.$$

Упражнение 21.5-1. Найдите $J_0\left(\frac{1}{2}\right)$ по интегральной формуле, используя шесть узлов.

§ 21.6. Метод общего оператора

В гл. 10 был описан общий метод (третий метод) нахождения формул, дающих точные значения для $1, x, x^2, \dots, x^N$. Аналогичным путем можно найти формулы, дающие точные значения для

$$\begin{aligned} 1, \cos \frac{\pi}{N} t, \cos \frac{2\pi}{N} t, \cos \frac{3\pi}{N} t, \dots, \cos \frac{(N-1)\pi}{N} t, \cos \pi t, \\ \sin \frac{\pi}{N} t, \sin \frac{2\pi}{N} t, \sin \frac{3\pi}{N} t, \dots, \sin \frac{(N-1)\pi}{N} t. \end{aligned} \quad (21.6-1)$$

Такие условия легко получить, если функция периодическая и используется $2N$ равноотстоящих узловых точек.

Пусть результат применения линейного оператора L к функции $f(t)$ выражается через значения функции в $2N$ равноотстоящих точках $t = 0, 1, 2, \dots, 2N-1$:

$$L[f(t)] = w_0 f(0) + w_1 f(1) + \dots + w_{2N-1} f(2N-1). \quad (21.6-2)$$

Определяющие уравнения для

$$\begin{aligned} 1, \cos \frac{\pi}{N} m, \cos \frac{2\pi}{N} m, \dots, \cos \frac{\pi(N-1)}{N} m, \cos \pi m, \\ \sin \frac{\pi}{N} m, \dots, \sin \frac{\pi(N-1)}{N} m \end{aligned}$$

Вследствие соотношений ортогональности (6.2-3)

$$\begin{aligned}\sum_{k=0}^{2N-1} \cos \frac{m\pi}{N} k \cos \frac{n\pi}{N} k &= N\delta_{mn} \quad (m, n \neq 0, N), \\ \sum_{k=0}^{2N-1} \sin \frac{m\pi}{N} k \cos \frac{n\pi}{N} k &= 0, \\ \sum_{k=0}^{2N-1} \sin \frac{m\pi}{N} k \sin \frac{n\pi}{N} k &= N\delta_{mn}, \\ \sum_{k=0}^{2N-1} 1 &= \sum_{k=0}^{2N-1} \cos^2 \pi k = 2N\end{aligned}$$

обратная матрица будет равна транспонированной матрице (21.6-3), кроме того, что первый и $(N-1)$ -й столбцы матрицы, соответствующие функциям 1 и $\cos tm$, будут разделены на $2N$:

$$\left(\frac{1}{N} \right) \begin{bmatrix} \frac{1}{2} & 1 & \dots & 1 \\ \frac{1}{2} & \cos \frac{\pi}{N} & \dots & \cos \frac{N-1}{N} \pi \\ \frac{1}{2} & \cos \frac{2\pi}{N} & \dots & \cos \frac{N-1}{N} 2\pi \\ \dots & \dots & \dots & \dots \\ \frac{1}{2} & \cos \frac{2N-1}{N} \pi & \dots & \cos \frac{(N-1)(2N-1)}{N} \pi \\ \frac{1}{2} & 0 & \dots & 0 \\ -\frac{1}{2} & \sin \frac{\pi}{N} & \dots & \sin \frac{N-1}{N} \pi \\ \frac{1}{2} & \sin \frac{2\pi}{N} & \dots & \sin \frac{N-1}{N} 2\pi \\ \dots & \dots & \dots & \dots \\ -\frac{1}{2} & \sin \frac{2N-1}{N} \pi & \dots & \sin \frac{(N-1)(2N-1)}{N} \pi \end{bmatrix}. \quad (21.6-4)$$

Применение этого третьего метода в общем случае можно рассмотреть на примере. Предположим, что нужно найти первую производную периодической функции в точке $t=0$. Для определенности положим, что исследуется случай $N=3$. Тогда составляющие «моментов» для 1, $\cos \frac{\pi}{3} t$, $\cos \frac{2\pi}{3} t$, $\cos t$, $\sin \frac{\pi}{3} t$, $\sin \frac{2\pi}{3} t$ будут равны

$$0, \quad 0, \quad 0, \quad 0, \quad \frac{\pi}{3}, \quad \frac{2\pi}{3}.$$

Умножив обратную матрицу на вектор моментов, получим

$$({}^{1/2})({}^{1/3}) \left\{ \begin{pmatrix} 1 & 2 & 2 & 1 & 0 & 0 \\ 1 & 1 & -1 & -1 & \sqrt{3} & \sqrt{3} \\ 1 & -1 & -1 & 1 & \sqrt{3} & -\sqrt{3} \\ 1 & -2 & 2 & -1 & 0 & 0 \\ 1 & -1 & -1 & 1 & -\sqrt{3} & \sqrt{3} \\ 1 & 1 & -1 & -1 & -\sqrt{3} & -\sqrt{3} \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \frac{\pi}{3} \\ \frac{2\pi}{3} \end{pmatrix} \right\} = ({}^{1/6}) \left\{ \begin{pmatrix} 0 \\ \sqrt{3}\pi \\ -\frac{\sqrt{3}}{3}\pi \\ 0 \\ \frac{\sqrt{3}}{3}\pi \\ -\sqrt{3}\pi \end{pmatrix} \right\}. \quad (21.6-5)$$

Результат представляет собой вектор весов; таким образом, искомая формула примет вид

$$f'(0) = \frac{\sqrt{3}\pi}{18} [3f(1) - f(2) + f(4) - 3f(5)].$$

Упражнения

21.6-1. Используя матрицу (21.6-5), найти формулу для $f''(0)$.

21.6-2. Используя (21.6-5), найдите формулу для вычисления

$$\int_0^{2N} \int_0^u f(t) dt du.$$

§ 21.7. Несколько замечаний относительно общего метода

В случае аппроксимации многочленами матричный метод был достаточно эффективен. Однако при аппроксимации рядом Фурье дело обстоит иначе. Различие состоит в том, что разложение функции в ряд Фурье находится легче, чем определяются веса при помощи умножения матриц (см. §§ 6.4 — 6.6). Если уже известны коэффициенты Фурье для данной функции, то, умножив их на «моменты», нетрудно получить окончательный результат. Таким образом, если известен ряд Фурье

$$f(t) = \frac{a_0}{2} + \sum_{k=1}^{N-1} a_k \cos \frac{\pi}{N} kt + \frac{a_N}{2} \cos \pi t + \sum_{k=1}^{N-1} b_k \sin \frac{\pi}{N} kt,$$

то, применив оператор $L(\cdot)$, получим

$$\begin{aligned} L[f(t)] &= \frac{a_0}{2} L(1) + \sum_{k=1}^{N-1} a_k L\left(\cos \frac{\pi}{N} kt\right) + \frac{a_N}{2} L(\cos \pi t) + \\ &+ \sum_{k=1}^{N-1} b_k L\left(\sin \frac{\pi}{N} kt\right) = \frac{a_0}{2} M_0 + \sum_{k=1}^{N-1} a_k M_k + \frac{a_N}{2} M_N + \sum_{k=1}^{N-1} b_k M_{N+k}. \end{aligned}$$

Коэффициенты ряда определяются независимо друг от друга и довольно просто, так как синусы и косинусы являются ортогональными функциями. Поэтому метод обратных матриц при аппроксимации рядами Фурье, хотя и представляет теоретический интерес, не имеет большого практического значения.

Как и в § 13.3, здесь необходимо указать особый случай, касающийся неопределенного интегрирования. Будет ли формула

$$y = \int_a^b f(x) dx$$

точна для $f(x) = 1, x, x^2, \dots$ или для $y = 1, x, x^2, \dots$ — не одно и то же. Как указывалось в § 13.3, при данной формуле точность для $f(x) = 1, x, x^2$ эквивалентна точности для $y = x, x^2, x^3, \dots$, поскольку интегрирование степеней x увеличивает степень на 1. Тогда мы нашли удобным добавить условие точности для $y = 1$ и таким образом восстановить эквивалентность обоих случаев.

В случае рядов Фурье, если сначала аппроксимировать подынтегральное выражение, а затем интеграл, то члены $\cos kx$ и $\sin kx$ перейдут друг в друга, а член 1 перейдет в x . Таким образом, в случае рядов Фурье эквивалентность приближения подынтегральной функции и интеграла будет потеряна. Поэтому с самого начала необходимо решить, какая часть задачи будет аппроксимироваться используемым множеством функций, так как аппроксимации на различных этапах решения задачи могут оказаться неэквивалентными. На практике обычно известны свойства лишь входных или выходных функций, и поэтому для аппроксимации выбирают те, свойства которых известны.

С эффектом неэквивалентности придется снова встретиться в гл. 26 при рассмотрении экспоненциальных функций. Так, при интегрировании функции $e^{a_i x}$ последняя переходит в саму себя (с точностью до постоянного множителя), за исключением случая $a_i = 0$.

ГЛАВА 22

СХОДИМОСТЬ РЯДОВ ФУРЬЕ

§ 22.1. Сходимость степенных рядов и рядов Фурье

Скорость сходимости степенных рядов и скорость, с которой сходится к функции последовательность множеств, совпадающих с ней в заданных точках, когда число этих точек возрастает, обе определяются расположением особенностей функции в комплексной плоскости. Это не значит, что, вообще говоря, поведение функции вдоль действ-

вительной оси не определяет скорости сходимости, а скорее означает, что трудно перейти от значений на действительной оси к скорости сходимости иначе, как через особенности. Особенности в комплексной плоскости часто практически невозможно получить; даже грубые оценки имеются редко. Например, рассматривается задача вычисления орбиты Луны. Как оценить положение ближайших особенностей в комплексной плоскости?

Напротив, сходимость рядов Фурье может быть легко определена по значениям функции вдоль действительной оси. В самом деле, между ними существуют простые соотношения. Цель настоящей главы вывести эти соотношения и показать, как ими пользоваться. Можно также оценить коэффициенты рядов Фурье до начала вычислений, что позволит сделать разумные оценки того, сколько членов понадобится для получения заданной точности (ср. с § 21.5). Таким образом, в стадии планирования вычислений метод Фурье имеет важные преимущества перед многочленным методом.

§ 22.2. Функции с простым разрывом

Начнем с примера (см. рис. 22.2-1) функции

$$y = \frac{t}{2\pi} \quad (-\pi < t < \pi), \quad (22.2-1)$$

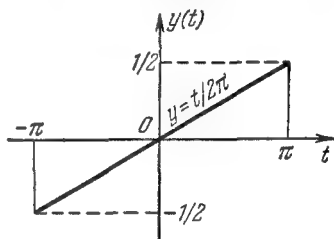


Рис. 22.2-1.

которая имеет один разрыв высоты 1 в точках $t = \pm \pi$ (предполагается, что $y(t)$ периодическая). Так как функция нечетная (т. е. $f(-t) = -f(t)$), то все члены с косинусами исчезают, в том числе и постоянный член. Коэффициенты при синусах определяются так:

$$\begin{aligned} b_k &= \frac{1}{\pi} \int_{-\pi}^{\pi} y(t) \sin kt dt = \frac{2}{\pi} \int_0^{\pi} \frac{t}{2\pi} \sin kt dt = \\ &= \frac{1}{\pi^2} t \frac{-\cos kt}{k} \Big|_0^{\pi} + \frac{1}{k\pi^2} \int_0^{\pi} \cos kt dt = \frac{\pi}{\pi^2 k} (-1) \cos \pi k = \frac{(-1)^{k-1}}{\pi k}. \end{aligned}$$

Следовательно,

$$y(t) = \frac{1}{\pi} \left(\sin t - \frac{\sin 2t}{2} + \frac{\sin 3t}{3} - \frac{\sin 4t}{4} + \dots \right). \quad (22.2-2)$$

Этот ряд сходится к $y(t)$ при всех значениях t , кроме точки разрыва ($t = \pm \pi$), где он, очевидно, сходится к нулю (среднему арифметическому двух значений $\frac{1}{2}$ и $-\frac{1}{2}$, которые являются предельными значениями при $t \rightarrow \pi$ или $-\pi$).

Если положить

$$t = t' - a + \pi,$$

то разрыв будет переведен в точку $t' = a$. Вместо (22.2-2) имеем (опуская штрих над t) для функции $y(t)$ со скачком в точке a :

$$\begin{aligned} y(t) &= \frac{1}{\pi} \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} \sin [(kt - ka) + k\pi] = \\ &= \frac{1}{\pi} \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} \sin (kt - ka) \cos k\pi = \frac{1}{\pi} \sum_{k=1}^{\infty} \frac{\sin (ka - kt)}{k} = \\ &= \frac{1}{\pi} \sum_{k=1}^{\infty} \frac{\sin ka}{k} \cos kt - \frac{1}{\pi} \sum_{k=1}^{\infty} \frac{\cos ka}{k} \sin kt. \end{aligned} \quad (22.2-3)$$

В этом ряде Фурье коэффициенты опять убывают как $\frac{1}{k}$.

Для функции с конечным числом M простых разрывов и линейной между разрывами, у которой скачок при $t = a_i$ имеет величину y_i , можно взять соответствующую линейную комбинацию рядов вида (22.2-3) (включая, быть может, еще член a_0), что дает

$$\begin{aligned} y(t) &= \frac{1}{\pi} \sum_{i=1}^M y_i \left(\sum_{k=1}^{\infty} \frac{\sin ka_i}{k} \cos kt - \sum_{k=1}^{\infty} \frac{\cos ka_i}{k} \sin kt \right) = \\ &= \frac{1}{\pi} \sum_{k=1}^{\infty} \left(\sum_{i=1}^M y_i \sin ka_i \right) \frac{\cos kt}{k} - \frac{1}{\pi} \sum_{k=1}^{\infty} \left(\sum_{i=1}^M y_i \cos ka_i \right) \frac{\sin kt}{k}, \end{aligned}$$

где мы формально переменили порядок суммирования.

Таким образом, функции с простыми разрывами и прямолинейными участками между разрывами дают ряды Фурье, коэффициенты которых убывают как $\frac{1}{k}$; кроме того, мы получили простой метод для построения таких рядов.

Для практических целей верно и обратное: если коэффициенты ряда убывают как $\frac{1}{k}$, то функция имеет простые разрывы. Часто встречается случай непрерывной *) ломаной, и потому его стоит исследовать отдельно. Вычислим сначала коэффициенты членов с ко-

*) Напоминаем, что для непрерывности необходимо также $f(\pi) = f(-\pi)$.

синусами, обозначая через t_i концы линейных участков. Коэффициенты при косинусах суть

$$\begin{aligned} \pi a_k &= \int_{-\pi}^{\pi} f(t) \cos ktdt = \sum_i \int_{t_i}^{t_{i+1}} f(t) \cos ktdt = \\ &= \sum_i \frac{1}{k} [f(t_{i+1}) \sin k(t_{i+1}) - f(t_i) \sin k(t_i)] - \sum_i \int_{t_i}^{t_{i+1}} \frac{f'(t) \sin kt dt}{k}. \end{aligned}$$

Когда суммирование выполняется на всем интервале, все члены уничтожаются, так как $f(t)$ предполагалась непрерывной. В интегральном члене $f'(t)$ является константой в каждом интервале, следовательно, результат интегрирования $\sin kt$ приводит еще к одному множителю $\frac{1}{k}$. В результате коэффициенты членов с косинусами убывают как $\frac{1}{k^2}$. То же самое происходит для членов с синусами. Таким образом, для непрерывной ломаной коэффициенты убывают как $\frac{1}{k^2}$.

Упражнение 22.2-1. Найти ряд Фурье для пилообразной функции

$$f(t) = \begin{cases} 1+t & \text{для } (-1 < t < 0), \\ t & \text{для } (0 < t < 1). \end{cases}$$

§ 22.3. Функция, имеющая непрерывные производные более высокого порядка

Если функция имеет непрерывную первую производную и непрерывную, кроме конечного числа точек, вторую производную, то можно поступить, как в § 22.2, интегрируя дважды, и показать, что коэффициенты убывают как $\frac{1}{k^3}$ *)

$$a_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \cos ktdt = -\frac{1}{\pi k} \int_{-\pi}^{\pi} f'(t) \sin ktdt,$$

так как из-за непрерывности проинтегрированный член обращается в нуль. Разобьем отрезок интегрирования и еще раз проинтегрируем

*) Более точное доказательство позволяет установить, что в этом случае коэффициенты Фурье стремятся к нулю как $\frac{1}{k^3}$. (Прим. ред.)

по частям

$$a_k = -\frac{1}{\pi k} \sum_{i=0}^M \int_{t_i}^{t_{i+1}} f'(t) \sin ktdt = -\frac{1}{\pi k^2} \sum_{i=0}^M \int_{-\pi}^{\pi} f''(t) \cos ktdt,$$

$$|a_k| \leq \frac{M_2}{\pi k^2} 2\pi,$$

где M_2 есть максимум второй производной. С b_k поступим аналогичным образом.

Если коэффициенты убывают уже как $\frac{1}{k^2}$, то можно заключить, что ряд Фурье сходится везде, так как

$$|a_k \cos kt + b_k \sin kt| \leq \frac{4M_2}{k^2}.$$

Та же схема, очевидно, применима к функциям, имеющим непрерывную*) m -ю производную; их коэффициенты Фурье убывают как $\frac{1}{k^{m+1}}$ (m — целое).

Упражнения

22.3-1. Найти ряд Фурье для

$$f(t) = \begin{cases} t(\pi - t) & \text{для } 0 \leq t \leq \pi, \\ -f(t) & \text{для } 0 > t \geq -\pi. \end{cases}$$

22.3-2. Пусть $f(t) = 1 - \frac{t^2}{\pi^2}$ ($-\pi \leq t \leq \pi$). Как ведут себя коэффициенты?

§ 22.4. Улучшение сходимости ряда Фурье

Разложение в ряд Фурье функций с разрывами или другими особенностями на действительной оси, особенно в концах интервала, часто встречается в практике. Так как уже известно, что особенности на действительной оси определяют скорость сходимости, естественно действовать почти так, как и в § 4.2: из данного ряда вычесть соответственно выбранный, также медленно сходящийся. Таким образом, для вычислений будут получаться быстрее сходящиеся ряды. Идея достаточно очевидна, и нет необходимости останавливаться на ней более подробно.

Коэффициенты a_k и b_k задаются интегралами, которые часто не могут быть вычислены и их приходится находить численно по значениям подынтегральной функции в нескольких узловых точках. Таким образом,

*) Коэффициенты Фурье имеют порядок $\frac{1}{k^{m+1}}$ также и для функций, у которых m -я производная кусочно непрерывна, т. е. может иметь конечное число конечных скачков. (Прим. ред.)

приходится снова сталкиваться с задачами мимикрии. Однако, прежде чем делать это, вычтем «канонические» разложения особенностей, такие как (22.2-3), а также интегралы от них, чтобы скомпенсировать особенности функции и ее производных. Тогда придем к численному нахождению коэффициентов быстро сходящегося ряда для очень гладкой функции; мимикрия не является уже для нее серьезной проблемой. Таким образом, вычитание особенностей не только уменьшает вычислительную работу, но также улучшает точность.

Упражнение 22.4-1. Вычислить первые три интеграла равенств (22.2-3) для получения канонической формы соответствующих особенностей.

§ 22.5. Спектр мощности

Как в чистой, так и в прикладной математике, обычно ищут инварианты представления — инварианты по отношению к классу преобразований. В классе периодических функций перенос осей

$$t = t' + b$$

не должен менять в представлении функции того, что не зависит от системы координат. Непосредственно видно, что коэффициенты Фурье a_k и b_k не обладают этим свойством и меняются при сдвиге оси, т. е. когда изменяется начало отсчета времени. Полагая $t = t' + b$ и используя периодичность $f(t)$, чтобы сдвинуть пределы в интеграле, получаем

$$\begin{aligned} a_k &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \cos ktdt = \frac{1}{\pi} \int_{-\pi}^{\pi} f(t' + b) \cos k(t' + b) dt' = \\ &= \left[\frac{1}{\pi} \int_{-\pi}^{\pi} f(t' + b) \cos kt' dt \right] \cos kb - \left[\frac{1}{\pi} \int_{-\pi}^{\pi} f(t' + b) \sin kt' dt \right] \sin kb = \\ &= a'_k \cos kb - b'_k \sin kb. \end{aligned}$$

Аналогично

$$b_k = a'_k \sin kb + b'_k \cos kb.$$

Хотя a_k и b_k не инвариантны, величина

$$\begin{aligned} a_k^2 + b_k^2 &= (a'_k \cos kb - b'_k \sin kb)^2 + (a'_k \sin kb + b'_k \cos kb)^2 = \\ &= (a'_k)^2 + (b'_k)^2, \end{aligned}$$

очевидно, инвариантна. Величину $a_k^2 + b_k^2$ называют *мощностью частоты k* и изображают в виде дискретного *спектра мощности*. Существует путаница относительно того, что дает *спектр $a_k^2 + b_k^2$* или $\sqrt{a_k^2 + b_k^2}$, но мы будем использовать слова «спектр мощности» и «спектр», имея в виду одну и ту же величину $a_k^2 + b_k^2$.

Упражнение 22.5-1. Вычислить спектр функции из упражнения 22.2-1.

§ 22.6. Явление Гиббса

Начнем с частного случая прямоугольной волны $H(t)$ с периодом 2π (рис. 22.6-1). Если вычислить сумму первых $2n$ членов, то все члены с косинусами будут равны нулю и получаем

$$H_{2n}(t) = \frac{1}{2} + \frac{2}{\pi} \sum_{k=1}^n \frac{1}{2k-1} \sin(2k-1)t. \quad (22.6-1)$$

Гиббс отметил, что частичная сумма H_{2n} превосходит функцию на некоторую величину (рис. 22.6-2). Более точно

$$H_{2n}\left(\frac{\pi}{2n}\right) \rightarrow 1,08949\dots, \text{ когда } n \rightarrow \infty. \quad (22.6-2)$$

Действительно, $H_{2n}(t)$ не только превосходит функцию $H(t)$, но и

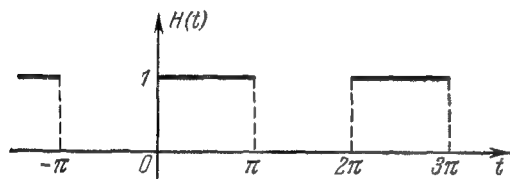


Рис. 22.6-1. Прямоугольная волна.

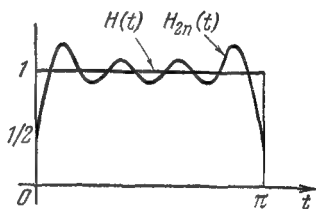


Рис. 22.6-2. Явление Гиббса.

имеет тенденцию колебаться около $H(t)$, и колебания уменьшаются медленно, когда t удаляется от разрыва.

Чтобы объяснить явление, запишем (22.6-1) как

$$\begin{aligned} H_{2n}(t) &= \frac{1}{2} + \frac{2}{\pi} \sum_{k=1}^n \int_0^t \cos(2k-1)x dx = \\ &= \frac{1}{2} + \frac{2}{\pi} \int_0^t \sum_{k=1}^n \cos(2k-1)x dx = \frac{1}{2} + \frac{1}{\pi} \int_0^t \frac{\sin 2nx}{\sin x} dx, \end{aligned} \quad (22.6-3)$$

где использована формула

$$\sum_{k=1}^n \cos(2k-1)x = \frac{\sin 2nx}{2 \sin x}.$$

Из (22.6-3) ясно, что максимум и минимум для $0 \leq t \leq \pi$ достигаются в точках

$$\frac{dH_{2n}(t)}{dt} = \frac{1}{\pi} \frac{\sin 2nt}{\sin t} = 0,$$

т. е. при

$$t = \frac{m\pi}{2n}, \quad m = 1, 2, \dots, 2n-1, \quad (22.6-4)$$

и что они чередуются. Их величины были вычислены Карслоу [7].

То, что верно для этой специальной функции, очевидно, верно и для более общих функций, так как разрыв можно рассматривать как возникающий из прямоугольной волны, прибавленной к главной функции.

Упражнение 22.6-1. Для функции в упражнении 22.2-1 начертить сумму первых четырех членов.

§ 22.7. Сигма-множители Ланцоша

Заменим быстро колеблющуюся функцию $H_{2n}(t)$ сглаженной функцией

$$\bar{H}_{2n}(t) = \frac{n}{\pi} \int_{t - \frac{\pi}{2n}}^{t + \frac{\pi}{2n}} H_{2n}(\tau) d\tau, \quad (22.7-1)$$

где усредняется одно полное колебание $H_{2n}(t)$, сосредоточенное около t .

Подставив (22.6-1) в (22.7-1), получаем

$$\begin{aligned} \bar{H}_{2n}(t) &= \frac{n}{\pi} \int_{t - \frac{\pi}{2n}}^{t + \frac{\pi}{2n}} \left[\frac{1}{2} + \frac{2}{\pi} \sum_{k=1}^n \frac{1}{2k-1} \sin(2k-1)\tau \right] d\tau = \\ &= \frac{n}{\pi} \left[\frac{\pi}{2n} + \frac{2}{\pi} \sum_{k=1}^n \frac{-1}{(2k-1)^2} \cos(2k-1)\tau \right]_{t - \frac{\pi}{2n}}^{t + \frac{\pi}{2n}} = \\ &= \frac{1}{2} + \frac{2}{\pi} \sum_{k=1}^n \frac{1}{2k-1} \cdot \frac{\sin\left[(2k-1)\frac{\pi}{2n}\right]}{(2k-1)\frac{\pi}{2n}} \sin(2k-1)t. \end{aligned} \quad (22.7-2)$$

Если сравнить это выражение с $H_{2n}(t)$, то получим дополнительный множитель

$$\sigma_{2k-1} = \frac{\sin(2k-1)\frac{\pi}{2n}}{(2k-1)\frac{\pi}{2n}} \quad (22.7-3)$$

для каждого члена при суммировании.

Эффект этого множителя σ_k состоит в том, чтобы уменьшить максимум с 0,08949 до 0,01187 и первый минимум с 0,04859 до 0,00473 и т. д. Таким образом, явление Гиббса сильно уменьшилось от присутствия σ -множителей, которые возникли из сглаживания $H_{2n}(t)$ на коротком интервале длиной π/n .

Мы изучили частный случай прямоугольной волны; покажем теперь, что влияние σ_k -множителей остается тем же для любого ряда Фурье.

Пусть $f(t)$ ($0 \leq t \leq 2\pi$) интегрируемая и пусть

$$a_k = \frac{1}{\pi} \int_0^{2\pi} f(t) \cos kt \, dt, \quad k = 0, 1, 2, \dots,$$

$$b_k = \frac{1}{\pi} \int_0^{2\pi} f(t) \sin kt \, dt, \quad k = 1, 2, \dots,$$

— коэффициенты Фурье. Тогда

$$f_n(t) = \frac{a_0}{2} + \sum_{k=1}^n (a_k \cos kt + b_k \sin kt).$$

Теперь вычислим

$$\begin{aligned} f_n(t) &= \frac{1}{\pi/n} \int_{t-\frac{\pi}{2n}}^{t+\frac{\pi}{2n}} f_n(\tau) \, d\tau = \frac{n}{\pi} \left[\frac{a_0}{2} \cdot \frac{\pi}{n} + \sum_{k=1}^n \left(a_k \frac{\sin k\tau}{k} - b_k \frac{\cos k\tau}{k} \right) \right] \Big|_{t-\frac{\pi}{2n}}^{t+\frac{\pi}{2n}} = \\ &= \frac{n}{\pi} \left(\frac{\pi}{2n} a_0 + \sum_{k=1}^n \left\{ \frac{a_k}{k} \left[\sin k \left(t + \frac{\pi}{2n} \right) - \sin k \left(t - \frac{\pi}{2n} \right) \right] - \right. \right. \\ &\quad \left. \left. - \frac{b_k}{k} \left[\cos k \left(t + \frac{\pi}{2n} \right) - \cos k \left(t - \frac{\pi}{2n} \right) \right] \right\} \right) = \\ &= \frac{a_0}{2} + \frac{n}{\pi} \sum_{k=1}^n \left(\frac{2a_k}{k} \sin \frac{\pi k}{2n} \cos kt + \frac{2b_k}{k} \sin \frac{\pi k}{2n} \sin kt \right) = \\ &= \frac{a_0}{2} + \sum_{k=1}^n \frac{\sin \frac{\pi k}{2n}}{\frac{\pi k}{2n}} (a_k \cos kt + b_k \sin kt) \end{aligned}$$

и опять получаем σ -множители (равенство (22.7-3))

$$\sigma_k = \frac{\sin \frac{\pi k}{2n}}{\frac{\pi k}{2n}}, \quad (22.7-4)$$

вставленные в различные коэффициенты ряда Фурье. Заметим, что члены $\sin kt$ и $\cos kt$ оба имеют множитель σ_k .

Упражнение 22.7-1. Применить σ -множители в уравнении 22.6-1 и начертить результат.

§ 22.8. Сравнение методов сходимости

Кроме метода σ -множителей Ландоша, существует хорошо известный метод Фейера, использующий средние арифметические частичных сумм. Метод Фейера совершенно исключает колебание, тогда как

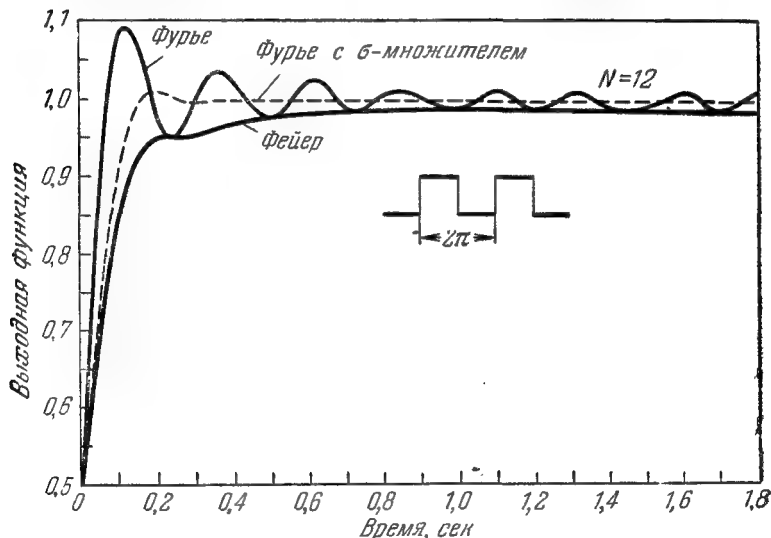


Рис. 22.8-1. Частичные суммы Фурье $= \frac{1}{2} + \frac{2}{\pi} \sum_{n=0}^N \frac{1}{2n+1} \sin(2n+1)t$;

Фейера $= \frac{1}{2} + \frac{1}{\pi(N+1)} \sum_{n=0}^N \frac{2(N-n)+1}{2n+1} \sin(2n+1)t$; Фурье с σ -мно-

жителем сходимости $= \frac{1}{2} + \frac{2}{\pi} \sum_{n=0}^N \frac{1}{2n+1} \frac{\sin \frac{(2n+1)\pi}{2N+2}}{\frac{(2n+1)\pi}{2N+2}} \sin(2n+1)t$.

метод Ландоша только сильно гасит его. Рис. 22.8-1 дает соответствующее сравнение кривых

ряда Фурье
суммы Фейера
 σ -множителя Ландоша

для двенадцатичленной аппроксимации прямоугольной волны. Преимущество метода σ -множителя очевидно.

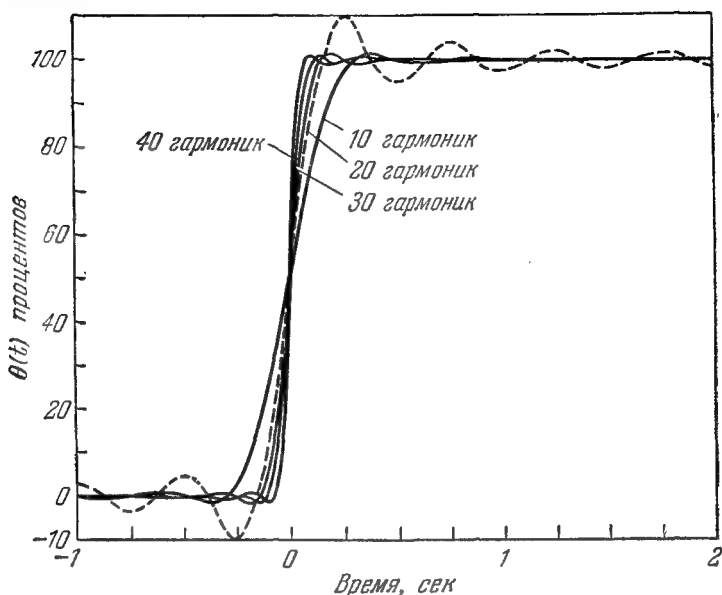


Рис. 22.8-2. Прямоугольная функция входа $f = \frac{1}{10} \frac{1}{\text{сек}}$; $T = 10 \text{ сек.}$
 Приближение Фурье и приближение Фурье с множителем сходимости.

Рис. 22.8-2 показывает скорость, с которой метод σ -множителя приближает прямоугольную волну как функцию n . Таким образом, даже для умеренного n «время роста» кривой очень невелико.

§ 22.9. Техника дифференцирования по Ланцошу

Иногда бывает необходимо продифференцировать ряд Фурье. При этом часто получаются нежелательные в физических задачах большие колебания высокой частоты.

Ланцош [23] предложил практический прием, состоящий в использовании

$$\frac{f\left(t + \frac{\pi}{n}\right) - f\left(t - \frac{\pi}{n}\right)}{\frac{2\pi}{n}} \quad (22.9-1)$$

в качестве оценки производной $f(t)$ в точке t , вместо предельного процесса, который имеет сомнительное физическое значение, для диф-

ференцирования приближения усеченным рядом Фурье. Заметим, что то же самое n возникает и в (22.9-1) и в порядке усеченного ряда. Мы не будем воспроизводить здесь выводы; заметим только, что *те же самые σ -множители* (22.7-4), введенные в формально дифференцированный ряд, дают соответствующий результат.

ГЛАВА 23

НЕПЕРИОДИЧЕСКИЕ ФУНКЦИИ. ИНТЕГРАЛ ФУРЬЕ

§ 23.1. Цель главы

В главах 21, 22 периодические функции аппроксимировались линейной комбинацией периодических — синусов и косинусов. Обратимся теперь к изучению непериодических функций и их аппроксимации синусами и косинусами. Вот пример непериодической функции, составленной из двух периодических,

$$y(t) = \cos t + \cos(\sqrt{2}t)$$

(так как 1 и $\sqrt{2}$ несоизмеримы, то $y(t)$ непериодическая). Приближение непериодической функции периодическими ничуть не хуже, чем представление периодической функции рядом Тейлора, составленным из непериодических членов

$$1, x, x^2, \dots$$

Для представления периодической функции использовалось бесконечное, но дискретное множество частот. Теперь для представления непериодических функций придется брать все частоты или, может быть, все частоты из данного интервала*). Основным инструментом при этом служит интеграл Фурье, который будет сейчас коротко рассмотрен; за более строгим изложением отсылаем читателя к стандартным учебникам**).

Цель настоящей главы — обсудить (нестрогим образом) ряд вопросов, возникающих в вычислительной практике; именно:

1. Что происходит от того, что узлы выбираются равноотстоящими?
2. Каким образом удастся восстановить функцию с ограниченным спектром по ее узлам так, что можно интерполировать функцию для промежуточных значений аргумента или использовать ее в аналитической подстановке?

*) Функция, у которой все частоты находятся в некотором интервале, называется *функцией с ограниченным спектром*.

**) Одной из самых последних работ, в которой содержится введение в теорию интеграла Фурье, является [24].

3. Каков эффект того, что рассматривается лишь конечное число узлов для функции, существующей в бесконечном числе точек?

Это — те вопросы, на которые приближение многочленами, рассмотренное во второй части, практически не дает ответа, и они помогают выявить преимущества настоящего метода на этапах планирования, исполнения и обсуждения вычислений.

§ 23.2. Обозначения и краткое изложение результатов

Этот параграф посвящен введению обозначений и формулировкам некоторых результатов, которые будут получены, но не содержит доказательств большинства сделанных утверждений.

Ряд Фурье на интервале $-N \leq t \leq N$ можно записать так (см. (21.3-1) и (21.3-2)):

$$f(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos \frac{\pi}{N} kt + \sum_{k=1}^{\infty} b_k \sin \frac{\pi}{N} kt, \quad (23.2-1)$$

где

$$a_k = \frac{1}{N} \int_{-N}^N f(t) \cos \frac{\pi}{N} kt dt \quad (k=0, 1, 2, \dots), \quad (23.2-2)$$

$$b_k = \frac{1}{N} \int_{-N}^N f(t) \sin \frac{\pi}{N} kt dt \quad (k=1, 2, \dots).$$

Если использовать комплексное представление тригонометрических функций

$$\cos \theta = \frac{e^{i\theta} + e^{-i\theta}}{2}, \quad \sin \theta = \frac{e^{i\theta} - e^{-i\theta}}{2i},$$

то получим

$$f(t) = \sum_{k=-\infty}^{\infty} c_k e^{\frac{i\pi}{N} kt}, \quad (23.2-3)$$

где

$$c_k = \frac{1}{2N} \int_{-N}^N f(t) e^{-\frac{i\pi}{N} kt} dt. \quad (23.2-4)$$

Легко видеть, что

$$c_k = \begin{cases} \frac{a_k - ib_k}{2}, & k > 0, \\ \frac{a_0}{2}, & k = 0, \\ \frac{a_k + ib_k}{2}, & k < 0. \end{cases} \quad (23.2-5)$$

Комплексная форма ряда Фурье значительно удобнее в обращении при теоретических исследованиях, чем обычная, но вычисления, конечно, всегда проводятся с действительной формой (23.2-1) и (23.2-2). Простота формальных манипуляций с комплексной формой часто скрывает необходимость больших вычислений и в этом месте следует быть осторожным. Но сейчас не требуется проводить никаких реальных вычислений; нужно лишь понять, как влияют различные вычислительные методы на окончательный ответ. Для этой цели комплексная форма удобнее в понимании и в употреблении. Заметим, что в комплексной форме существуют и положительные и отрицательные частоты: для каждой положительной частоты мы заменили две функции, синус и косинус, единой экспоненциальной, но имеющей как положительную, так и отрицательную частоту.

Покажем сперва (§ 23.3), что соответственно представлению рядом Фурье ((23.2-3) и (23.2-4)) периодической функции имеется представление интегралом Фурье любой функции

$$f(t) = \int_{-\infty}^{\infty} F(\sigma) e^{2\pi i \sigma t} d\sigma, \quad (23.2-6)$$

где

$$F(\sigma) = \int_{-\infty}^{\infty} f(t) e^{-2\pi i \sigma t} dt. \quad (23.2-7)$$

Функция $F(\sigma)$, грубо говоря, соответствует коэффициентам c_k в ряде Фурье (23.2-3) и (23.2-4).

Это — *спектральная функция* или *спектральная плотность*; $F(\sigma)$ описывает амплитуду частоты σ в функции $f(t)$. Ее называют также *преобразованием Фурье* функции $f(t)$. Обычно употребляют большие и маленькие буквы для обозначения преобразования Фурье и соответствующей функции. Изменение знака показателя степени у экспоненты — единственное отличие между этими двумя функциями. Чтобы различать переменную преобразования, мы используем латинские и греческие буквы, латинские — для обозначения времени, греческие — для обозначения частот.

Как мы видели в § 21.2, эффект выборки должен вызывать наложение различных частот. Это никоим образом не связано с периодичностью разлагаемой функции, а только со смешением частот, использованных для ее представления. Это наложение и является причиной того, что понятие функции с ограниченным спектром играет ведущую роль в теории; если все частоты ограничены интервалом

$$-\Omega < \sigma < \Omega$$

ширины 2Ω и если интервал между узлами равен Δt , то, чтобы избежать наложения, необходимо иметь

$$2\Omega \Delta t < 1. \quad (23.2-8)$$

Иначе говоря, чтобы избежать наложения, мы должны иметь по крайней мере два узла на самой короткой волне, присутствующей в разложении.

Естественно возникает вопрос: «можно ли восстановить функцию с ограниченным спектром по ее узлам, если в самом деле следовать (23.2-8)»? Интуитивным аргументом за то, что это возможно, является аналогия с интерполяционной схемой Лагранжа (§ 8.3). Функция

$$\frac{\sin \pi t}{\pi t}$$

обладает тем свойством, что ее значение равно 1 при $x=0$ и нулю при $t=\pm 1, \pm 2, \pm 3, \dots$. Таким образом, функция

$$\frac{\sin \pi (t-k)}{\pi (t-k)} \quad (23.2-9)$$

играет ту же роль, что и L_k в интерполяционной схеме Лагранжа, так как она равна 1 при $t=k$ и нулю во всех остальных узлах. Далее,

$$f(k) \frac{\sin \pi (t-k)}{\pi (t-k)}$$

принимает значение функции $f(k)$ при $t=k$ и нуль во всех остальных узлах; следовательно, формальное разложение

$$\sum_{k=-\infty}^{\infty} f(k) \frac{\sin \pi (t-k)}{\pi (t-k)}$$

проходит через все узлы. Легко видеть, что

$$\frac{\sin \pi t}{\pi t} = \int_{-1}^1 \frac{1}{2} e^{i\pi t \sigma} d\sigma. \quad (23.2-10)$$

Следовательно, согласно (23.2-6) частоты, присутствующие в $\frac{\sin \pi t}{\pi t}$, ограничены. Здесь σ меняется от -1 до 1 ; чтобы представить экспоненту в форме (23.2-6), следует взять

$$\frac{\sin \pi t}{\pi t} = \int_{-1/2}^{1/2} e^{2\pi i \sigma' t} d\sigma',$$

после чего получаем интервал $-\frac{1}{2} < \sigma' < \frac{1}{2}$. Это соответствует ожиданиям; взяв интервал выборки $\Delta t=1$, мы получили, что частота пробегает интервал от $-\frac{1}{2}$ до $\frac{1}{2}$. По существу, полученное разло-

жение представляет собой *теорему выборки* теории информации*); если заданы необходимые равноотстоящие узлы для функции с ограниченным спектром, то по этим узлам можно восстановить функцию.

По аналогии с § 22.5 величина

$$|F(\sigma)|^2 \quad (23.2-11)$$

часто называется спектром мощности; слово «мощность» пришло из инженерных применений, и мы используем его просто по традиции. Спектр мощности, или просто спектр, похож на обычный оптический спектр: как входящий луч света разлагается призмой на отдельные цвета (частоты), так интеграл Фурье разлагает функцию времени $f(t)$ на гармоники амплитуды $F(\sigma)$.

В § 23.7 будет показано, что для функции, заданной на бесконечной прямой ($-\infty < t < \infty$), выбор узлов на ограниченном отрезке должен расширить спектр, и чем короче длина участка, на котором берутся данные, тем больше расширение спектра.

Сказанное выше раскрывает суть этой главы и показывает ее значение для вычислительной практики. Ясно, что и эффект расстановки узлов, и эффект выбора конечного числа узлов могут быть хотя бы отчасти поняты в рамках теории функций с ограниченным спектром, но их нельзя понять в рамках классической теории полиномиальных приближений.

В оставшейся части главы развивается теория интеграла Фурье в той мере, в какой это необходимо, чтобы немного пояснить сказанное выше. Мы не хотим оказаться вовлеченными в вопросы математической строгости, так как класс функций, которые действительно пробуют аппроксимировать на вычислительных машинах, ограничен «функциями с хорошим поведением» и не включает патологические случаи, которыми так часто занимаются математики.

Для тех, кому трудно манипулировать с комплексными числами, можно предложить следующие простые упражнения для практики. Если они вызовут затруднения, будет лучше, прежде чем двинуться дальше, еще раз повторить эту тему.

Упражнения

23.2-1. Доказать, что $|e^{ix}| = 1$ (для действительных x).

23.2-2. Доказать, что тригонометрические формулы сложения следуют из равенства $e^{i\alpha}e^{i\beta} = e^{i(\alpha+\beta)}$.

23.2-3. $\frac{1}{a+ib} = u+iv$; выразить u , v через a и b .

*) В нашей литературе по теории информации для этой теоремы принято название «теорема Котельникова». Соответствующее разложение называют рядом Котельникова. См. В. А. Котельников, «О пропускной способности «эфира» и «провода»». Материалы к Первому всесоюзному съезду по вопросам реконструкции дела связи, М., 1933. (Прим. ред.)

23.2-4. $u + iv = (a + ib)^{\frac{1}{2}}$, найти u, v (обратить внимание на знак u).

23.2-5. Доказать, что если $c_k = c_{-k}$, то сумма ряда $\sum_{k=-\infty}^{\infty} c_k e^{ikt}$ действительна.

23.2-6. Доказать, что если m и k — целые, то

$$\int_{-\pi}^{\pi} e^{imx} e^{-ikx} dx = \begin{cases} 0, & m \neq k, \\ 2\pi, & m = k. \end{cases}$$

23.2-7. Пусть $f(t) = t$ ($-\pi < t < \pi$). Найти коэффициенты Фурье в разложении

$$f(t) = \sum_{k=-\infty}^{\infty} c_k e^{ikt}.$$

23.2-8. Пусть $f(t) = \int_{-\infty}^{\infty} F(\sigma) e^{2\pi i \sigma t} d\sigma$.

Найти преобразование Фурье для $f'(t)$, $f''(t)$, $f^{(k)}(t)$.

23.2-9. Пусть

$$f(t) = \begin{cases} t & (-\pi < t < \pi), \\ 0 & \text{в остальных точках.} \end{cases}$$

Найти $F(\sigma)$.

23.2-10. Пусть

$$F(\sigma) = \begin{cases} \sigma & (-\pi < \sigma < \pi), \\ 0 & \text{в остальных точках.} \end{cases}$$

Найти $F(t)$.

§ 23.3. Интеграл Фурье

Чтобы «вывести» интеграл Фурье из ряда Фурье, исключим коэффициенты c_k , подставляя (23.2-4) в (23.2-3):

$$f(t) = \sum_{k=-\infty}^{\infty} \left[\int_{-N}^N f(t) e^{-\frac{i\pi}{N} kt} dt \right] e^{\frac{i\pi}{N} kt} \frac{1}{2N}.$$

Предполагается, что функция $f(t)$ периодическая в интервале $-N \leq t \leq N$. Для приближения непериодической функции заставим N стремиться к бесконечности, для чего положим

$$\frac{1}{2N} = \Delta\sigma,$$

и заметим, что в этой сумме расстояние между точками, в которых берутся соседние экспоненты, ведет себя как $\Delta\sigma$, которая при $N \rightarrow \infty$ будет стремиться к нулю. Имеем

$$f(t) = \sum_{k=-\infty}^{\infty} \left[\int_{-N}^N f(t) e^{-2\pi i k t \Delta\sigma} dt \right] e^{2\pi i k t \Delta\sigma} \Delta\sigma.$$

Но $k\Delta\sigma \rightarrow \sigma$ и при $N \rightarrow \infty$, получаем интеграл

$$f(t) = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} f(t) e^{-2\pi i t \sigma} dt \right] e^{2\pi i \sigma t} d\sigma.$$

Положив теперь

$$F(\sigma) = \int_{-\infty}^{\infty} f(t) e^{-2\pi i t \sigma} dt, \quad (23.3-1)$$

получаем

$$f(t) = \int_{-\infty}^{\infty} F(\sigma) e^{2\pi i \sigma t} d\sigma. \quad (23.3-2)$$

Говорят, что функция $F(\sigma)$ является преобразованием Фурье функции $f(t)$. Соотношения между этими функциями почти симметричны, разница только в знаке показателя $\sigma \rightarrow -\sigma$. Обе функции несут одну и ту же информацию, так как каждая может быть найдена из другой, но только в разных формах: $f(t)$ в области времени, а $F(\sigma)$ в области частот. Возможностью изучения информации в любом из двух видов и объясняется в основном ценность представления функции в виде интеграла Фурье.

Вышесказанное не является, конечно, строгим доказательством, это всего лишь довод в пользу того, что высказанные утверждения верны. Устремляя N к бесконечности, мы отказываемся от предположений относительно периодичности функции.

§ 23.4. Преобразование Фурье некоторых функций

Соотношения (23.3-1) и (23.3-2) показывают, что каждой функции $f(t)$ соответствует преобразование $F(\sigma)$ и наоборот. Существуют обширные таблицы ([6], [8]) формул, связывающих $f(t)$ и $F(\sigma)$. Выведем только некоторые из них — те, которые понадобятся для дальнейшей работы.

В качестве первого примера рассмотрим функцию с ограниченным спектром

$$f(t) = \int_{-\infty}^{\infty} F(\sigma) e^{2\pi i \sigma t} d\sigma, \quad (23.4-1)$$

где (рис. 23.4-1)

$$F(\sigma) = \begin{cases} \frac{1}{2\Omega} & \text{для } |\sigma| < \Omega, \\ 0 & \text{для } |\sigma| > \Omega. \end{cases} \quad (23.4-2)$$

Таким образом, $F(\sigma)$ ограничивает единичную площадь, а $f(t)$ имеет все частоты в интервале $-\Omega < \sigma < \Omega$ и никаких вне его. Используя (23.4-2), запишем (23.4-1) в виде

$$\begin{aligned} f(t) &= \frac{1}{2\Omega} \int_{-\Omega}^{\Omega} e^{2\pi i \sigma t} d\sigma = \\ &= \frac{1}{2\Omega} \cdot \frac{e^{2\pi i \sigma t}}{2\pi i t} \Big|_{-\Omega}^{\Omega} = \\ &= \frac{e^{2\pi i \Omega t} - e^{-2\pi i \Omega t}}{2i} \cdot \frac{1}{2\Omega \pi t} = \frac{\sin 2\pi \Omega t}{2\pi \Omega t}, \end{aligned} \quad (23.4-3)$$

что соответствует (23.2-10) (рис. 23.4-2). Используя (23.3-1), замечаем, что преобразование имеет вид

$$F(\sigma) = \int_{-\infty}^{\infty} \frac{\sin 2\pi \Omega t}{2\pi \Omega t} e^{-2\pi i \sigma t} dt, \quad (23.4-4)$$

где $F(\sigma)$ дана в (23.4-2).

В качестве второй иллюстрации соотношения между преобразованиями Фурье предположим, что задана пара функций $f(t)$, $F(\sigma)$, причем

$$f(t) = \int_{-\infty}^{\infty} F(\sigma) e^{2\pi i \sigma t} d\sigma. \quad (23.4-5)$$

Какой функции $f_1(t)$ соответствует преобразование Фурье $F(\sigma) e^{2\pi i \sigma y}$?

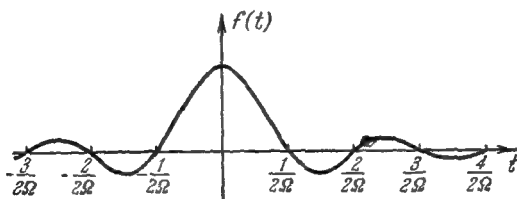


Рис. 23.4-2. Функция с ограниченным спектром $\left(\frac{\sin 2\pi \Omega t}{2\pi \Omega t} \right)$.

Имеем

$$f_1(t) = \int_{-\infty}^{\infty} F(\sigma) e^{2\pi i \sigma y} e^{2\pi i \sigma t} d\sigma = \int_{-\infty}^{\infty} F(\sigma) e^{2\pi i \sigma (y+t)} d\sigma.$$

Положим $y+t=z$; тогда из (23.4-5) следует

$$f_1(t) = f(z) = f(y+t) = \int_{-\infty}^{\infty} [F(\sigma) e^{2\pi i \sigma y}] e^{2\pi i \sigma t} d\sigma. \quad (23.4-6)$$

Таким образом, в результате умножения функций на экспоненту аргумент преобразования сдвигается. Уравнение (23.4-6) иногда называют *теоремой сдвига*.

§ 23.5. Функции с ограниченным спектром и теорема выборки

В § 23.2 был дан наглядный, но не очень строгий вывод *теоремы выборки*. Вследствие ее важности для вычислительной практики дадим другой ее вывод.

Центральная идея теоремы выборки состоит в том, чтобы брать узловые точки функции с ограниченным спектром вдвое чаще наивысшей частоты. Понятие функции с ограниченным спектром широко используется во многих областях науки. На практике часто встречается случай, когда вне некоторого интервала $F(\omega)$ очень мала, и модель функции с ограниченным спектром является полезной. Однако необходимо сказать несколько слов предостережения. Если функция с ограниченным спектром, то математическая модель говорит, что она не может быть «ограничена во времени», т. е. не может быть равна нулю для всех $|t| > t_0$ ни при каком t_0 . В частности, если $f(t)$ представляет электрический ток, то придется допустить, что он шел всегда и будет идти вечно. Соответственно если $f(t)$ «ограничена во времени», то она не может быть функцией с ограниченным спектром. Очевидно, не нужно считать математическую модель реально существующей; она является полезной аппроксимацией физических явлений, но не обязательно точно соответствует им.

Частота наложения ясно связана со смещением частот из-за того, что узлы выбраны равноотстоящими. Чтобы вывести теорему выборки, начнем с формул

$$F(\sigma) = \int_{-\infty}^{\infty} f(t) e^{-2\pi i \sigma t} dt, \quad f(t) = \int_{-\infty}^{\infty} F(\sigma) e^{2\pi i \sigma t} d\sigma. \quad (23.5-1)$$

Если $F_1(\sigma)$ периодическая (рис. 23.5-1, а) с периодом $-\Omega < \sigma < \Omega$, то $F_1(\sigma)$ имеет разложение в ряд Фурье

$$F_1(\sigma) = \sum_{k=-\infty}^{\infty} c_k e^{\frac{i\pi}{\Omega} k \sigma}, \quad (23.5-2)$$

где

$$c_k = \frac{1}{2\Omega} \int_{-\Omega}^{\Omega} F_1(\sigma) e^{-\frac{i\pi}{\Omega} k \sigma} d\sigma = \frac{1}{2\Omega} f_1\left(\frac{k}{2\Omega}\right) = \frac{1}{2\Omega} f\left(\frac{k}{2\Omega}\right). \quad (23.5-3)$$

Теперь рассмотрим прямоугольный импульс $P(\sigma)$ (рис. 23.5-1, б)

$$P(\sigma) = \begin{cases} \frac{1}{2\Omega} & \text{для } |\sigma| < \Omega, \\ 0 & \text{для } |\sigma| > \Omega. \end{cases}$$

По (23.4-3) его преобразование есть

$$p(t) = \frac{\sin 2\pi\Omega t}{2\pi\Omega t}. \quad (23.5-4)$$

Наконец, поскольку предполагалось, что $F(\sigma)$ имеет спектр, ограниченный интервалом $-\Omega < \sigma < \Omega$, рассмотрим

$$F(\sigma) = F_1(\sigma) P(\sigma) 2\Omega$$

и используем (23.5-2) и (23.5-3):

$$F(\sigma) = \sum_{k=-\infty}^{\infty} c_k e^{\frac{\pi i}{\Omega} k \sigma} P(\sigma) 2\Omega = \sum_{k=-\infty}^{\infty} f_1\left(\frac{k}{2\Omega}\right) P(\sigma) e^{\frac{\pi i}{\Omega} k \sigma}.$$

Теперь произведем обратное преобразование и применим теорему сдвига (равенство (23.4-6)):

$$f(t) = \sum_{k=-\infty}^{\infty} f_1\left(\frac{k}{2\Omega}\right) \frac{\sin 2\pi\Omega\left(t - \frac{k}{2\Omega}\right)}{2\pi\Omega\left(t - \frac{k}{2\Omega}\right)} = \sum_{k=-\infty}^{\infty} f\left(\frac{k}{2\Omega}\right) \frac{\sin \pi(2\Omega t - k)}{\pi(2\Omega t - k)}.$$

(25.5-5)

Таким образом, получилась теорема выборки.

Для самой частоты наложения невозможно восстановить функцию, используя теорему выборки, так как если $\Delta t = s$ и $f(t) = \sin \pi t$

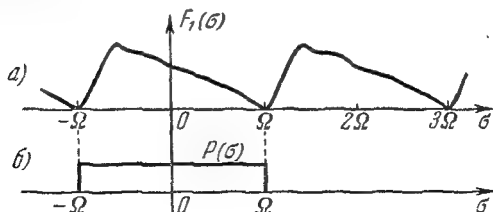


Рис. 23.5-1. а) $F_1(\sigma)$ — периодическая; б) $P(\sigma)$ — прямоугольный импульс.

то мы получим все значения в узлах равными нулю и по теореме выборки $f(t) \equiv 0$.

§ 23.6. Теорема свертки

Другим полезным соотношением, связанным с преобразованием Фурье, является *теорема свертки*. Предположим, что имеются две функции $f(t)$ и $g(t)$. Свертка $f(t)$ с $g(t)$ определяется как

$$h(t) = \int_{-\infty}^{\infty} f(s) g(t-s) ds. \quad (23.6-1)$$

Заметим, что если для фиксированного t написать $t-s-s'$, то получим

$$h(t) = \int_{-\infty}^{\infty} f(t-s') g(s') ds', \quad (23.6-2)$$

а следовательно, свертка f с g та же, что и свертка g с f .

Поставим теперь вопрос: «Что является преобразованием Фурье для свертки $h(t)$ »? По определению имеем

$$H(\sigma) = \int_{-\infty}^{\infty} h(t) e^{-2\pi i t \sigma} dt.$$

Используя (23.6-1), получим

$$\begin{aligned} H(\sigma) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(s) g(t-s) ds e^{-2\pi i t \sigma} dt = \\ &= \int_{-\infty}^{\infty} f(s) e^{-2\pi i s \sigma} \left[\int_{-\infty}^{\infty} g(t-s) e^{-2\pi i (t-s) \sigma} dt \right] ds = \\ &= \int_{-\infty}^{\infty} f(s) e^{-2\pi i s \sigma} G(\sigma) ds = F(\sigma) G(\sigma). \end{aligned} \quad (23.6-3)$$

Таким образом, *преобразование свертки двух функций есть произведение их преобразований*. Вышеприведенный формальный вывод доказывает утверждение для двух функций времени; в упражнении 23.6-1 рассматривается случай свертки двух функций от частоты σ .

Интересный результат, который следует из теоремы свертки, касается свертки $f(t)$ с $f(-t)$

$$h(t) = \int_{-\infty}^{\infty} f(s) f(-t+s) ds.$$

Это равно преобразованию произведения преобразований

$$h(t) = \int_{-\infty}^{\infty} F(\sigma) F(-\sigma) e^{2\pi i \sigma t} d\sigma.$$

Полагая $t=0$, имеем

$$h(0) = \int_{-\infty}^{\infty} f^2(s) ds = \int_{-\infty}^{\infty} F(\sigma) F(-\sigma) d\sigma. \quad (23.6-4)$$

Упражнения

23.6-1. Доказать, что если

$$H(\sigma) = \int_{-\infty}^{\infty} F(\tau) G(\sigma - \tau) d\tau,$$

то

$$h(t) = f(t) g(t).$$

23.6-2. Доказать формулу, соответствующую (23.6-4), используя упражнение 23.6-1.

§ 23.7. Эффект конечного суммирования

Мы можем рассматривать конечное число узловых точек, выбранное из бесконечной последовательности узловых точек, как произведение истинной функции времени и прямоугольного импульса

$$p(t) = \begin{cases} 1 & |t| < T, \\ 0 & |t| > T. \end{cases}$$

Таким образом, имеем

$$f_1(t) = p(t) f(t)$$

и по теореме свертки (§ 23.6)

$$F_1(\sigma) = \int_{-\infty}^{\infty} F(\tau) \frac{\sin 2\pi T(\tau - \sigma)}{2\pi T(\tau - \sigma)} d\tau. \quad (23.7-1)$$

Чтобы понять смысл этого выражения, рассмотрим функцию $f(t) = e^{2\pi i \sigma_1 t}$. Спектр $f(t)$ состоит из единственной частоты $\sigma = \sigma_1$ и свертка представляет собой «размазывание» этой частоты в спектр

$$F_1(\sigma) = \frac{\sin 2\pi T(\sigma_1 - \sigma)}{2\pi T(\sigma_1 - \sigma)}. \quad (23.7-2)$$

Чем больше T , тем уже центральный пик (рис. 23.4-2). В некотором смысле это размазывание и представляет эффект конечной выборки, когда мы пытаемся определить частоту по конечному куску чистой синусоиды. Если теперь представить спектр состоящим из многих частот, то каждая из них будет размазана тем же самым множителем (23.7-2); для предельного непрерывного распределения получим (23.7-1). В оптике это размазывание соответствует разрешающей способности: чем длиннее по времени сигнал, тем лучше можно разделить две близкие частоты.

ГЛАВА 24

ЛИНЕЙНЫЕ ФИЛЬТРЫ. СГЛАЖИВАНИЕ
И ДИФФЕРЕНЦИРОВАНИЕ

§ 24.1. Введение

Слово «фильтр», как и выражение «спектр мощности», возникли из физических явлений, изучаемых теорией электрических цепей, но получили значительно более широкое применение. Фильтры, с которыми мы будем иметь дело, удаляют из функции некоторые частоты, примерно так, как масляный фильтр удаляет из жидкости частицы определенных размеров. Теория построения фильтров очень развита, но мы можем здесь лишь коснуться этой темы; придется удовлетвориться демонстрацией основной идеи и способов расчета некоторых фильтров. Тех, кто интересуется построением фильтров для конкретных инженерных ситуаций, а не вычислителей, мы отсылаем к стандартным курсам теории электрических цепей и проектирования фильтров.

В этой главе будет изучена роль фильтров в двух типичных ситуациях: с целью сглаживания и дифференцирования данных. Мы старательно избегали этих тем во второй части, в крайнем случае коротко упоминали о них; обе они являются трудными и деликатными, и приближение многочленами не дает о них никакого представления. Дальнейшие примеры будут даны в гл. 25.

Типичный способ использования многочленных приближений для дифференцирования состоит в том, чтобы провести интерполяционный многочлен через узловые точки, продифференцировать этот многочлен, а затем вычислить производную в узловых точках. Достаточно посмотреть на остаточный член интерполяционного многочлена (см. (8.6-1))

$$\frac{(x - x_1)(x - x_2) \dots (x - x_{n+1}) f^{(n+1)}(\theta)}{(n+1)!},$$

чтобы увидеть, что в узловых точках аппроксимирующий многочлен почти наверняка пересекает функцию; следовательно, узловая точка есть почти наихудшее возможное место для вычисления производной.

Иногда, прежде чем дифференцировать, для получения гладкого многочлена используют метод наименьших квадратов. Этот процесс минимизирует сумму квадратов отклонений от первоначальных данных в узловых точках, но ничего не говорит о поведении между узловыми точками. Таким образом, хотя можно надеяться, что производная оценена правильно, нельзя быть слишком уверенным, что многочлен, проведенный по способу наименьших квадратов, не имеет

между узловыми точками колебаний, которые будут сильно влиять на оценку производной.

Использование для этой цели функций с ограниченным спектром, хотя и не отвечает на все вопросы, все же дает некоторое понимание ситуации.

§ 24.2. Пример простого сглаживающего фильтра

Пусть дана функция $f(t)$, зависящая от времени, и по той или иной причине требуется сгладить равноотстоящие узлы f_k по линейной формуле

$$\frac{f_k + f_{k+1} + f_{k+2}}{3} \equiv g_k \quad (t_k = hk). \quad (24.2-1)$$

Что сделается при этом со спектром функции?

Предположим сначала, что $f(t)$ — простая синусоида $f(t) = e^{2\pi i \sigma t}$ (σ — действительное число). Удобно ввести обозначение

$$2\pi\sigma = \omega.$$

Здесь σ измеряется в циклах в единицу времени, а ω — в радианах в единицу времени. Таким образом,

$$f(t) = e^{i\omega t}.$$

Возьмем для удобства $h = \Delta t = 1$; тогда, используя (24.2-1), получим

$$f_k = e^{i\omega k}, \quad g_k = \frac{e^{i\omega(k+1)}}{3} (e^{-i\omega} + 1 + e^{i\omega}) = \frac{e^{i\omega(k+1)}}{3} (1 + 2 \cos \omega).$$

Следовательно, значения f умножились на величину, не зависящую от k ,

$$\omega = \left| \frac{1 + 2 \cos \omega}{3} \right|.$$

Важно заметить, что множитель получился не зависящим от k вследствие линейности g_k . Если бы g_k были нелинейными функциями от f_k , то этот результат не был бы справедлив.

В спектре множитель становится равным

$$\left(\frac{1 + 2 \cos \omega}{3} \right)^2, \quad (24.2-2)$$

что равно нулю, когда $\omega = \frac{2\pi}{3}$ или $\sigma = \frac{1}{3}$. Если начертить эту кривую (рис. 24.2-1) как функцию частоты σ , то можно убедиться, что

формула сглаживания (24.2-1) дает сильный эффект подавления большинства частот в верхней половине спектра. Таким образом, метод функции с ограниченным спектром показывает, что делает с первоначальным сигналом простой линейный фильтр (уравнение (24.2-1)): он взвешивает частоты согласно кривой на рис. 24.2-1.

Мы задали вопрос о действии линейного фильтра на единственную частоту, но на самом деле увидели его действие на каждую частоту; следовательно, используя представление преобразованием Фурье, мы можем работать с функциями, составленными из всех частот.

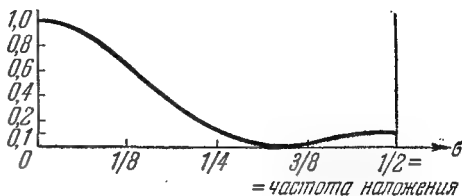


Рис. 24.2-1. Фильтры $1/3$ ($f_k + f_{k+1} + f_{k+2}$).

§ 24.3. Пример построения фильтра

Как известно, расстояние между узлами непосредственно связано с частотой наложения в спектре. Предположим, что функция (сигнал) имеет ограниченный спектр и максимальную частоту 10 циклов в секунду. Это вынуждает ставить 20 точек в секунду, чтобы избежать смешивания частот.

Допустим, что интерес представляют лишь частоты, меньшие 5 циклов в секунду. Мы не рискуем взять 10 точек в секунду, так как тогда часть спектра с частотами выше 5 циклов в секунду наложилась бы на интересующий нас интервал.

Фильтр § 24.2 дает возможность устранить верхнюю половину спектра. Как только фильтр установлен, он уменьшает эти частоты по крайней мере в девять раз.

Мы поступим еще лучше, поставив вслед за первым второй сглаживающий фильтр

$$\frac{g_k + g_{k+1} + g_{k+2} + g_{k+3}}{4} \equiv h_k. \quad (24.3-1)$$

На простую синусоиду $e^{2\pi i \sigma t} = e^{i\omega t}$ этот фильтр действует так:

$$\begin{aligned} h_k &= \frac{e^{i\omega k}}{4} (1 + e^{i\omega} + e^{2i\omega} + e^{3i\omega}) = \frac{e^{i\omega k}}{4} \cdot \frac{1 - e^{4i\omega}}{1 - e^{i\omega}} = \\ &= \frac{e^{i\omega(k+2)}}{4e^{i\omega/2}} \cdot \frac{e^{2i\omega} - e^{-2i\omega}}{e^{i\omega/2} - e^{-i\omega/2}} = \frac{e^{i(k+5/2)}}{4} \cdot \frac{\sin 2\omega}{\sin \frac{\omega}{2}}. \end{aligned}$$

Следовательно, спектр умножается на величину

$$\left(\frac{\sin 2\omega}{4 \sin \frac{\omega}{2}} \right)^2 = \left(\frac{\sin 4\pi\sigma}{4 \sin \pi\sigma} \right)^2, \quad (24.3-2)$$

которая имеет нули при $\sigma = 1/4, 1/2, \dots$

Результат последовательного действия двух фильтров (24.2-1) и (24.3-1) представляет собой фильтр

$$h_k = \frac{f_k + 2f_{k+1} + 3f_{k+2} + 3f_{k+3} + 2f_{k+4} + f_{k+5}}{12}, \quad (24.3-3)$$

который очень хорошо подавляет верхнюю половину спектра (см. рис. 24.3-1, где нарисован логарифм спектра в зависимости от σ).

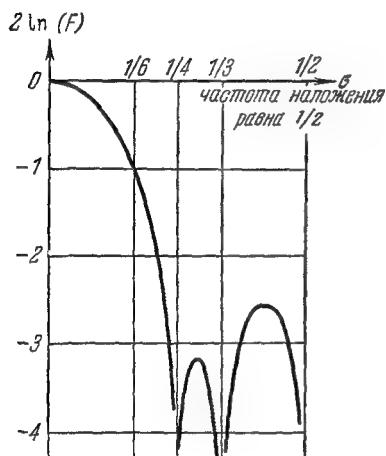


Рис. 24.3-1. Частотная характеристика фильтра

$$h_k = \frac{1}{12} (f_k + 2f_{k+1} + 3f_{k+2} + 3f_{k+3} + 2f_{k+4} + f_{k+5}).$$

После такого сглаживания можно спокойно опустить через один члены h_k , сохраняя, скажем, h_{2k} , и пользоваться результатом без серьезных опасений относительно смещения (частота наложения равна теперь 5 сек^{-1}), если только мощность верхней части спектра не была слишком велика по сравнению с нижней. При окончательном использовании результатов необходимо учитывать эффект, полученный вследствие фильтрации на частотах, больших чем 5 циклов в секунду. Таким образом, действие фильтра (24.3-3) позволяет выбирать часть значений функции, не имея при этом неприятностей смещения. Точность для частот между $1/3$ и $1/2$ первоначальной частоты наложения будет не очень хорошей.

Упражнение 24.3-1. Рассмотреть влияние сглаживания на спектр сигнала при усреднении m последовательных членов.

$$\text{О т в е т: } \left(\frac{\sin \pi m \sigma}{m \sin \pi \sigma} \right)^2.$$

§ 24.4. Фильтры вообще

Из сказанного выше вытекает, что какую-нибудь линейную комбинацию равноотстоящих узлов можно рассматривать как некоторую «фильтрацию» функции. В прошлом, пока соответствующая теория не была достаточно ясна, нередко случалось, что некоторые невинно

выглядевшие преобразования, выполненные на ранней стадии вычислений, существенно влияли на полученный результат, который затем интерпретировался как физический эффект, а не как эффект обработки данных.

В качестве примера представим себе белый шум (в котором, как и в белом свете, присутствуют все частоты с одинаковой амплитудой, но случайной фазой) и при обработке его удалим почти все частоты, кроме небольшого интервала; тогда последующий анализ показал бы наличие преобладающей частоты, и одна и та же частота появлялась бы независимо от источника белого шума, будь то рыночные цены или количество пятен на Солнце. Такие данные могут иметь спектр, отличный от гладкого; но анализ данных, испорченных шумом, требует значительных обсуждений.

С другой стороны, как было показано в § 24.3, разумное использование фильтров может быть очень полезным. Эта область обширна и обычные курсы теории цепей, возможно, являются лучшим источником дальнейших сведений, так как эти идеи редко просачиваются в вычислительные круги. Дальнейшие рекомендации, авторитетные, но трудные для чтения, можно найти у Блекмана и Тьюки *).

§ 24.5. Анализ простых формул для дифференцирования

Возможно, самой простой и наиболее широко используемой формулой для вычисления производной через равноотстоящие узлы является формула

$$f'_k = \frac{f_{k+1} - f_{k-1}}{2h} \quad (h — шаг). \quad (24.5-1)$$

Проанализируем, что она делает с различными частотами. Предположим, что

$$f(t) = e^{2\pi i \omega t} = e^{i\omega t}. \quad (24.5-2)$$

Пусть в k -й узловой точке $t_k = kh$. Тогда

$$f_k = e^{2\pi i \omega kh} = e^{i\omega kh}.$$

Дифференцирование $f(t)$ дает

$$\left. \frac{df(t)}{dt} \right|_{t=kh} = \frac{df_k}{dt} = i\omega e^{i\omega kh}. \quad (24.5-3)$$

С другой стороны, (24.5-1) дает оценку

$$f'_k = e^{i\omega kh} \frac{e^{i\omega h} - e^{-i\omega h}}{2h} = i\omega e^{i\omega kh} \frac{\sin \omega h}{\omega h}. \quad (24.5-4)$$

*). См. [2].

Отношение вычисленного ответа (24.5-4) к верному (24.5-3) опять не зависит от h и равно

$$\frac{\sin \omega h}{\omega h}$$

(см. рис. 23.4-2). Таким образом, формула (24.5-1) явно уменьшает амплитуду всех частот, кроме $\omega = 0$. Как можно было бы ожидать, при $\omega = \frac{\pi}{h}$, т. е. $\sigma = \frac{1}{2h}$, для производной получается нулевое значение ($\sigma = \frac{1}{2h}$ есть частота наложения).

Подобный эффект имеет место и для второй производной

$$f_k'' = \frac{f_{k+1} - 2f_k + f_{k-1}}{h^2}.$$

Используя (24.5-2), получаем оценку

$$f_k'' = e^{i\omega kh} \frac{e^{i\omega h} - 2 + e^{-i\omega h}}{h^2} = -\omega^2 e^{i\omega kh} \frac{2(1 - \cos \omega h)}{\omega^2 h^2} = -\omega^2 e^{i\omega kh} \frac{\sin^2 \frac{\omega h}{2}}{\left(\frac{\omega h}{2}\right)^2}.$$

Правильный ответ есть

$$f_k'' = -\omega^2 e^{i\omega kh},$$

так что отношение результатов равно

$$\frac{\sin^2 \frac{\omega h}{2}}{\left(\frac{\omega h}{2}\right)^2} \sim 1 - \frac{\omega^2 h^2}{12} + \dots,$$

которое опять уменьшает все величины, кроме $\omega = 0$. Для частоты $h\omega = \pi$ находим значение производной, равное $-\frac{4}{h^2} e^{i\pi k}$.

§ 24.6. Как избежать вычисления производных?

Часто можно избежать задачи оценки производной. Для примера предположим, что имеются значения f_k и известно, что теоретически $f(t)$ удовлетворяет дифференциальному уравнению второго порядка вида

$$f'' = H(f, t).$$

Используя его, можно перейти от значений $f(t)$ к значениям $f''(t)$ и проинтегрировать последнюю функцию, чтобы получить $f'(t)$. Часто интегрирование предпочитают дифференцированию при условии, что отрезок, на котором имеются данные, не настолько велик, чтобы не-

большие систематические погрешности интегрирования дали в результате большую ошибку.

Для частных случаев известны многочисленные другие хитрости, но, по-видимому, общей теории, когда и как можно избежать дифференцирования, пока нет.

§ 24.7. Метод Филон

Те же самые простые идеи, которыми мы пользовались выше, могут дать представление о том, что в действительности делает формула, не вдаваясь в детальный анализ, который был произведен для сглаживающего и дифференцирующего фильтров. В то же время будут выведены некоторые полезные формулы.

Часто встречается задача вычисления интегралов вида

$$I(k) = \int_a^b f(t) \cos kt \, dt. \quad (24.7-1)$$

Филон *) предложил метод, который, формально говоря, относится ко второй части этой книги, так как основывается на многочленной аппроксимации $f(t)$. Точнее, интервал (a, b) делится на $2N$ интервалов, и в каждом двойном интервале $f(t)$ аппроксимируется квадратным трехчленом. Таким образом, это напоминает составную формулу Симпсона, кроме того, что под интегралом имеется еще множитель $\cos kt$.

Идеи метода очень просты, но требуют длинных выкладок **), и мы приведем лишь результаты. Возьмем шаг

$$h = \frac{b-a}{2N}, \quad (24.7-2)$$

и пусть C_{2n} — сумма всех четных ординат $f(t) \cos kt$, за исключением того, что от первой и последней ординат берется половина

$$C_{2n} = \frac{1}{2} f(a) \cos ka + f(a+h) \cos k(a+h) + \\ + f(a+2h) \cos k(a+2h) + \dots + \frac{1}{2} f(b) \cos kb. \quad (24.7-3)$$

Пусть также C_{2n-1} — сумма всех нечетных ординат

$$C_{2n-1} = f(a+h) \cos k(a+h) + f(a+3h) \cos k(a+3h) + \dots \\ \dots + f(b-h) \cos k(b-h). \quad (24.7-4)$$

*) L. N. G. Filon, Proc. Roy. Soc. Edm., XLIX (1928—1929).

**) См., например, [40].

Тогда

$$\int_a^b f(t) \cos kt \, dt = h \{ \alpha [f(b) \sin kb - f(a) \sin ka] + \beta C_{2n} + \gamma C_{2n-1} \}, \quad (24.7-5)$$

где

$$\begin{aligned} \alpha &= \frac{\theta^2 + \theta \sin \theta \cos \theta - 2 \sin^2 \theta}{\theta^3} = \frac{2\theta^3}{45} - \frac{2\theta^5}{315} + \frac{2\theta^7}{4725} + \dots, \\ \beta &= \frac{2 [\theta (1 + \cos^2 \theta) - 2 \sin \theta \cos \theta]}{\theta^3} = \frac{2}{3} + \frac{2\theta^2}{15} + \frac{4\theta^4}{105} + \frac{2\theta^6}{567} + \dots, \\ \gamma &= \frac{4 (\sin \theta - \theta \cos \theta)}{\theta^3} = \frac{4}{3} - \frac{2\theta^2}{15} + \frac{\theta^4}{210} - \frac{\theta^6}{11340} + \dots \end{aligned} \quad (24.7-6)$$

и

$$\theta = kh = \frac{k(b-a)}{2N}. \quad (24.7-7)$$

Такая же формула относится к интегралу

$$\int_a^b f(t) \sin kt \, dt = h \{ -\alpha [f(b) \cos kb - f(a) \cos ka] + \beta S_{2n} + \gamma S_{2n-1} \}, \quad (24.7-8)$$

где S_{2n} и S_{2n-1} — соответствующие суммы для $f(t) \sin kt$.

Попробуем понять эти формулы при помощи простого применения идей, о которых шла речь выше.

Величины, которые здесь вычисляются, являются коэффициентами Фурье, так что ищется амплитуда гармоники $f(t)$ с частотой k (радиан). Величина h есть расстояние между узлами, а величина

$$\theta = kh$$

— его произведение на исследуемую частоту. Теорема выборки показывает, что нельзя надеяться надежно вычислить амплитуду, если не выбраны по крайней мере два узла на участке длиной $\frac{2\pi}{k}$; но даже тогда потребовалось бы бесконечное множество узлов. Так как интервал (a, b) конечен, то можно ожидать, что θ много меньше, чем π (этот вопрос будет еще обсуждаться в следующей главе).

Таблицы коэффициентов α , β , γ даны Трантером [40] (для θ в радианах) до $\theta = 1,50$ и Копалом ([20], стр. 539) (для θ в градусах) до 45° . В обоих случаях и в первоначальной работе Филона ничего не сказано, почему таблицы обрываются именно в этом месте. Почему, если $f(t)$ ведет себя как многочлен, нельзя применять формулу так далеко, как хотелось бы, и делать k , а следовательно и θ , совсем большим? Раз сделана «аналитическая замена» функции многочленом, то интегрирование делается аналитически. Что останавливает

нас? Очевидно, в основе этого лежит теорема выборки, определяющая величину θ , которой еще можно пользоваться.

Мы также ожидаем, что при $\theta \rightarrow 0$ получается формула Симпсона. Уравнения (24.7-5) с (24.7-6) и (24.7-7) ясно показывают это, причем отклонения будут порядка θ^2 .

В принципе можно произвести детальный анализ и определить точно, что делает метод Филона с каждой частотой, но мы этим заниматься не будем. Метод позволяет сделать это, но выкладки слишком запутаны, чтобы воспроизводить их в учебнике.

Упражнения

24.7-1. Вывести равенство (24.7-5).

24.7-2. Вывести равенство (24.7-8).

§ 24.8. Заключительные замечания

Мы не изложили (и не собираемся делать этого) задачи сглаживания и дифференцирования*) во всех деталях. Как уже было сказано, это очень деликатные вопросы. Однако основные нити соответствующих рассуждений должны быть ясны. Нам дана функция (сигнал), искаженная шумом, например сигнал отдаленной планеты, или функция, выданная электронной машиной с ошибками округления. Пусть даже у нас есть теоретические оценки спектра функции и шума. Тогда мы сталкиваемся с трудным вопросом создания фильтра, который сделал бы с сигналом то, что нам нужно, а также, если можно, удалял бы ненужный шум.

На практике часто приходится сталкиваться с задачей оценки амплитуд нескольких частот из суммарного спектра, который задан. Бывает, что измеренный спектр выглядит примерно как на рис. 24.8-1.

Этот рисунок наводит на мысль о том, что мы видим здесь прямоугольник**) белого шума (наверное, вместе с частотами, наложившимися на интервал, допускаемый данной выборкой) плюс сигнал на низких частотах. Мы могли бы поэтому спланировать обрезающий фильтр, который удалил бы верхнюю часть спектра (вместе с некоторой частью сигнала), и объединить его со вторым фильтром, который делал бы с сигналом то, что нам нужно.

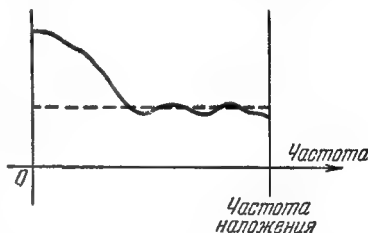


Рис. 24.8-1. Типичный спектр.

*) Напоминаем читателю σ -метод Ландоша из § 22.9.

**) В гл. 32 дается спектральный анализ шума, получающегося от флуктуаций выборки; что же касается соответствующего изучения спектра шума округления, то его еще нужно сделать.

Следует понимать, что, прежде чем строить фильтр, нужно представлять себе, из каких частот составлен данный сигнал. К сожалению, вопрос об измерении спектра мощности (§ 24.4) лежит за пределами начального курса, и мы должны поэтому оставить здесь эти две важные темы, лишь слегка коснувшись их.

ГЛАВА 25

ИНТЕГРАЛЫ И ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ

§ 25.1. Содержание главы

В этой главе будут рассмотрены некоторые формулы интегрирования с точки зрения их влияния на различные частоты, а также формулы для интегрирования дифференциальных уравнений. В конечном счете, речь будет идти о методах интегрирования, основанных на приближении функциями с ограниченным спектром. В частности, будут рассмотрены методы Чебышева.

Мы видели, что использование обозначения $e^{2\pi i \sigma t}$ полезно при различных преобразованиях; оно удобно также при оценке точности формул. Это последнее обстоятельство уже неявно использовалось; исследуем его более внимательно.

Предположим, что рассматривается синусоида $A \sin 2\pi \sigma t$. Если измерять качество формулы величиной ошибки, то, чтобы установить, насколько данная ошибка серьезнее, нужно различать случаи

$$A=1 \text{ и } A=1000.$$

Более разумно оценивать относительную ошибку. Однако для значений вблизи начала координат относительная ошибка может быть очень большой, и здесь следует предпочесть использование абсолютной ошибки.

При рассмотрении комплексной синусоиды $Ae^{2\pi i \sigma t}$ для действительных σ и t мы видим, что ее модуль является фиксированным по величине

$$|Ae^{2\pi i \sigma t}| = |A|$$

и относительная ошибка дает удовлетворительный критерий точности.

Кроме того, при вычислении модуля величины $Ae^{2\pi i \sigma t + i\varphi}$ фазовый угол φ исчезает. Таким образом, при изучении функции $Ae^{2\pi i \sigma t}$ часто можно пренебречь фазовым углом φ или ввести фазовый угол в коэффициент A (который может быть комплексным числом). Но, как будет видно в дальнейшем, относительной фазой различных функций пренебрегать нельзя.

§ 25.2. Метод передаточной функции для интегрирования

В гл. 13 рассматривался класс формул для вычисления неопределенных интегралов. Основное выражение (13.3-1) имело вид

$$y_{n+1} = a_0 y_n + a_1 y_{n-1} + a_2 y_{n-2} + h(b_{-1} y'_{n+1} + b_0 y'_n + b_1 y'_{n-1} + b_2 y'_{n-2}), \quad (25.2-1)$$

где y'_k — значения подынтегральной функции, а y_k — вычисленные значения

$$y(t) = y(0) + \int_0^t y'(x) dx.$$

Рассмотрим эти формулы еще раз, в свете теперешнего подхода. Предположим сначала, что подынтегральная функция (вход) есть синусоида

$$y'(t) = A_1 e^{2\pi i \sigma t} = A_1 e^{i \omega t} \quad (\omega = 2\pi \sigma). \quad (25.2-2)$$

(Индекс 1 относится ко входу.) Истинное значение таково:

$$y(t) = -\frac{i A_1}{\omega} e^{i \omega t} + C, \quad C = y(0) + \frac{i A_1}{\omega}. \quad (25.2-3)$$

Так как основная формула (25.2-1) линейна, то можно полагать, что вычисленные значения (выход) имеют ту же частоту, хотя, возможно, другие фазу и амплитуду. Пусть A_0 — амплитуда выхода, причем величину A_0 можно считать комплексной, чтобы включить в нее фазовый угол. Таким образом, предполагается, что вычисленная $y(t)$ с точностью до ошибок округления имеет вид

$$y(t) = A_0 e^{i \omega t}. \quad (25.2-4)$$

Подставим теперь эти функции (25.2-2) и (25.2-4) в основную формулу (25.2-1) и разрешим ее относительно отношения $\frac{A_0}{A_1}$:

$$\frac{A_0}{A_1} = \frac{h(b_{-1} + b_0 e^{-i \omega h} + b_1 e^{-2i \omega h} + b_2 e^{-3i \omega h})}{1 - a_0 e^{-i \omega h} - a_1 e^{-2i \omega h} - a_2 e^{-3i \omega h}}. \quad (25.2-5)$$

Это отношение, называемое *передаточной функцией*, является множителем, на который умножается амплитуда данной частоты, для получения выходной амплитуды той же частоты. Идея передаточной функции есть просто формальное утверждение того, что часто делалось в гл. 24 при рассмотрении поведения отдельной частоты.

Рассмотрим более детально простой случай правила трапеций (12.2-1)

$$y_{n+1} = y_n + \frac{h}{2} (y'_{n+1} + y'_n).$$

Очевидно, в (25.2-1)

$$\begin{aligned} a_0 &= 1, & a_1 &= a_2 = 0, \\ b_{-1} &= b_0 = \frac{1}{2}, & b_1 &= b_2 = 0, \end{aligned}$$

и передаточная функция имеет вид

$$\frac{A_0}{A_1} = \frac{h}{2} \cdot \frac{1 + e^{-i\omega h}}{1 - e^{-i\omega h}} = \frac{h \cos \frac{\omega h}{2}}{2i \sin \frac{\omega h}{2}} = -i \left(\frac{h}{2} \operatorname{ctg} \frac{\omega h}{2} \right). \quad (25.2-6)$$

Сравнивая фазовые углы этого результата с истинным значением (25.2-3), видим, что они чисто мнимые, и мы имеем правильную фазу для всех ω . Если сравнить амплитуды, оставляя в стороне начальные условия как несущественные, то следует сравнить $\frac{h}{2} \operatorname{ctg} \frac{\omega h}{2}$ с $\frac{1}{\omega}$. Для маленьких ωh имеем

$$\frac{h}{2} \operatorname{ctg} \frac{\omega h}{2} = \frac{h}{2} \left[\frac{2}{\omega h} - \frac{\omega h}{6} - \frac{(\omega h)^3}{360} - \dots \right] = \frac{1}{\omega} - \frac{\omega h^2}{12} - \frac{\omega^3 h^4}{720} - \dots$$

Следовательно, отношение вычисленного ответа к точному, равное

$$1 - \frac{\omega^2 h^2}{12} - \frac{\omega^4 h^4}{720} - \dots,$$

отличается от 1 на величину порядка $-\frac{\omega^2 h^2}{12}$.

И передаточная функция, и отношение вычисленного ответа к точному являются функциями частоты ω и шага h . Отношение содержит только произведение ωh и показывает, как выбор шага влияет на частоту при одной и той же ошибке.

Теперь рассмотрим формулу Симпсона (12.2-2)

$$y_{n+1} = y_{n-1} + \frac{h}{3} (y'_{n+1} + 4y'_n + y'_{n-1}).$$

Для (25.2-1) здесь имеем

$$\begin{aligned} a_0 &= 0, & a_1 &= 1, & a_2 &= 0, \\ b_{-1} &= \frac{1}{3}, & b_0 &= \frac{4}{3}, & b_1 &= \frac{1}{3}, & b_2 &= 0. \end{aligned}$$

Передаточная функция имеет вид

$$\frac{A_0}{A_1} = \frac{h(1 + 4ie^{-i\omega h} + e^{-2i\omega h})}{3(1 - e^{-2i\omega h})} = -\frac{ih}{3} \cdot \frac{\cos \omega h + 2}{\sin \omega h} \quad (25.2-7)$$

и опять дает правильную фазу для всех ω (множитель $-i$) и для маленьких ωh амплитуду порядка $\frac{1}{\omega}$.

Наконец, рассмотрим правило трех восьмых (12.2-3). Передаточная функция здесь выглядит так:

$$\frac{3h}{8} \cdot \frac{1 + 3e^{-i\omega h} + 3e^{-2i\omega h} + e^{-3i\omega h}}{1 - e^{-3i\omega h}} = (-i) \frac{3h}{8} \frac{\cos \frac{3\omega h}{2} + 3 \cos \frac{\omega h}{2}}{\sin \frac{3\omega h}{2}}. \quad (25.2-8)$$

Чтобы стандартизовать приведенные формулы, возьмем отношение вычисленного по ним ответа к точному и выберем масштаб ω так, чтобы h стало равным единице.

Тогда получим:

для правила трапеций

$$H_1(\omega) = \frac{\omega}{2} \operatorname{ctg} \frac{\omega}{2};$$

для формулы Симпсона

$$H_2(\omega) = \frac{\omega}{3} \frac{2 + \cos \omega}{\sin \omega}; \quad (25.2-9)$$

для правила трех восьмых

$$H_3(\omega) = \frac{3\omega}{8} \cdot \frac{\cos \frac{3\omega}{2} + 3 \cos \frac{\omega}{2}}{\sin \frac{3\omega}{2}}.$$

Соответствующие графики представлены на рис. 25.2-1. По оси ординат использована логарифмическая шкала, ибо если будет допущена ошибка в определении этого отношения, скажем, вдвое, то не имеет значения — в какую сторону. По оси абсцисс отложены как

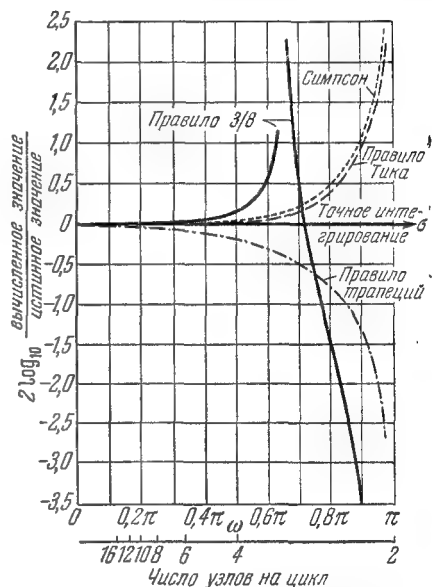


Рис. 25.2-1. Частотная характеристика некоторых формул интегрирования.

значения частоты ω , так и значения скорости выборки $\frac{2\pi}{\omega}$ (оси направлены в противоположные стороны). Легко видеть, что, если судить по величине ошибки, формула Симпсона — лучшая из трех (для достаточно большой частоты выборки).

На рис. 25.2-1 показана также кривая для формулы

$$y_{n+1} = y_{n-1} + h(0,3584y'_{n+1} + 1,2832y'_n + 0,3584y'_{n-1}).$$

Коэффициенты ее были определены экспериментально Лео Тиком так, чтобы в интервале $0 \leq \omega \leq \frac{\pi}{2}$ ошибка имела чебышевскую форму (см. гл. 19), а также чтобы формула была точной для $\omega = 0$. Ясно,

что скорость выборки была здесь удвоенной по сравнению с Найквистом.

Тот факт, что график для правила трапеций идет вниз, тогда как для формулы Симпсона — вверх, требует некоторого пояснения. Формула Симпсона повышает амплитуду высоких частот, тогда как правило трапеций стремится сгладить их. Хотя в действительности мало что известно об эффектах округления, неожиданные скачки функции, обусловленные округлением, порождают высокие частоты; следовательно, формула Симпсона увеличивает их, тогда как правило трапеций стремится их сгладить. Если рассматривать формулу Симпсона как функцию скорости выборки, то при пяти узлах на цикл ошибка на шаг составляет около $1,5\%$, что почти всегда слишком много. При семи узлах на цикл получается грубое, но иногда полезное приближение, а при десяти совершается ошибка, меньшая 0,001, что во многих случаях достаточно хорошо.

До сих пор рассматривалась главным образом единственная частота $\omega = 2\pi\sigma$. Предположим, что берется произвольная функция (вход)

$$f_1(t) = \int_{-\infty}^{\infty} F(\sigma) e^{2\pi i \sigma t} d\sigma.$$

Каждый член преобразуется при помощи передаточной функции

$$G(\omega) = G(2\pi\sigma) = G_1(\sigma)$$

и на выходе получается

$$f_0(t) = \int_{-\infty}^{\infty} F(\sigma) G_1(\sigma) e^{2\pi i \sigma t} d\sigma.$$

Справа имеем преобразование произведения и, применяя теорему свертки (§ 23.6), получаем

$$f_0(t) = \int_{-\infty}^{\infty} f_1(\tau) g_1(t - \tau) d\tau. \quad (25.2-10)$$

Таким образом, свертка преобразования $g_1(t)$ передаточной функции $g_1(\sigma)$ с входной функцией $f_1(t)$ дает выходную функцию $f_0(t)$.

Передаточная функция $G_1(\sigma)$ содержит всю информацию о формуле интегрирования и в некотором смысле эквивалентна применявшейся формуле.

Упражнение 25.2-1. Рассмотреть условия, необходимые для того, чтобы формула интегрирования имела правильную фазу ($-t$).

§ 25.3. Общие формулы интегрирования

Применение передаточной функции дает точную фазу для трех формул, которые исследовались в § 25.2, а также для формул, исследованных в гл. 24. Когда исследуется общая формула (25.2-1) для случаев, перечисленных в таблице 13.7-1, дело будет обстоять не так гладко. Например, при использовании метода Адамса — Башфорта получаем ($h=1$)

$$\frac{9 + 19e^{-i\omega} - 5e^{-2i\omega} + e^{-3i\omega}}{24(1 - e^{-i\omega})} = -i \frac{[9e^{\frac{i\omega}{2}} + 19e^{-\frac{i\omega}{2}} - 5e^{-\frac{3i\omega}{2}} + e^{-\frac{5i\omega}{2}}]}{48 \sin \frac{\omega}{2}}$$

и числитель не является чисто мнимым, так как мнимая часть выражения в скобках для маленьких ω есть

$$\left[9 \sin \frac{\omega}{2} - 19 \sin \frac{\omega}{2} + \right. \\ \left. + 5 \sin \frac{3\omega}{2} - \sin \frac{5\omega}{2} \right] \sim \omega^3.$$

Действительная часть этого выражения имеет вид

$$\left[9 \cos \frac{\omega}{2} + 19 \cos \frac{\omega}{2} - \right. \\ \left. - 5 \cos \frac{3\omega}{2} + \cos \frac{5\omega}{2} \right] \sim 24,$$

так что отклонение передаточного числа от чисто мнимого очень мало для маленьких ω .

То, что фаза неправильна, не следует принимать слишком всерьез. Хотя и амплитуду, и фазу можно изучать отдельно, здесь интересно лишь, насколько полученный ответ отличается от истинного. Рассмотрим поэтому величину

$$\left| \frac{i\omega A_0}{A_1} - 1 \right|, \quad (25.3-1)$$

которая дает хорошую единую меру ошибки*). Отношение $\left| \frac{\omega A_0}{A_1} \right|$,

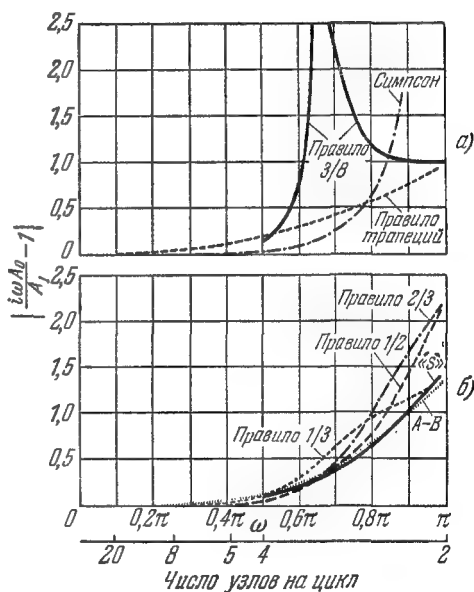


Рис. 25.3-1.

*) В некоторых задачах, например во многих задачах акустики, ошибка в фазе имеет второстепенное значение; но в других задачах, таких, как определение взаимных спектров, они могут иметь первостепенную важность.

которое использовалось раньше, может быть близко к 1 и все же иметь большую ошибку в фазе; мы не встретились с этим до сих пор, так как в изучавшихся случаях фаза была в точности правильной.

Рис. 25.3-1 показывает эту величину как функцию частоты

Т а б л и ц а 25.3-1

$\left| \frac{i\omega A_0}{A_f} - 1 \right|$ для различных методов интегрирования

Частота выборки	Угол	Трапеций	Адамса — Башфорта	Симпсона	Грех восьмых
25,00	0,08π	$5,27 \times 10^{-2}$	$1,05 \times 10^{-4}$	$2,23 \times 10^{-5}$	$5,06 \times 10^{-5}$
16,67	0,12π	$1,19 \times 10^{-2}$	$5,29 \times 10^{-4}$	$1,14 \times 10^{-4}$	$2,61 \times 10^{-4}$
12,50	0,16π	$2,11 \times 10^{-2}$	$1,66 \times 10^{-3}$	$3,66 \times 10^{-4}$	$8,49 \times 10^{-4}$
10,00	0,20π	$3,31 \times 10^{-2}$	$4,02 \times 10^{-3}$	$9,08 \times 10^{-4}$	$2,15 \times 10^{-3}$
8,33	0,24π	$4,78 \times 10^{-2}$	$8,25 \times 10^{-3}$	$1,92 \times 10^{-3}$	$4,67 \times 10^{-3}$
7,14	0,28π	$6,53 \times 10^{-2}$	$1,51 \times 10^{-2}$	$3,66 \times 10^{-3}$	$9,16 \times 10^{-3}$
6,25	0,32π	$8,57 \times 10^{-2}$	$2,54 \times 10^{-2}$	$6,44 \times 10^{-3}$	$1,68 \times 10^{-2}$

Частота выборки	Угол	Одной трети	Половины	Двух третей	«S»
25,00	0,08π	$5,05 \times 10^{-5}$	$5,00 \times 10^{-5}$	$3,44 \times 10^{-5}$	$1,28 \times 10^{-4}$
16,67	0,12π	$2,60 \times 10^{-4}$	$2,55 \times 10^{-4}$	$1,76 \times 10^{-4}$	$6,24 \times 10^{-4}$
12,50	0,16π	$8,39 \times 10^{-4}$	$8,10 \times 10^{-4}$	$5,66 \times 10^{-4}$	$1,88 \times 10^{-3}$
10,00	0,20π	$2,11 \times 10^{-3}$	$2,00 \times 10^{-3}$	$1,41 \times 10^{-3}$	$4,34 \times 10^{-3}$
8,33	0,24π	$4,53 \times 10^{-3}$	$4,18 \times 10^{-3}$	$3,02 \times 10^{-3}$	$8,51 \times 10^{-3}$
7,14	0,28π	$8,75 \times 10^{-3}$	$7,85 \times 10^{-3}$	$5,78 \times 10^{-3}$	$1,49 \times 10^{-2}$
6,25	0,32π	$1,57 \times 10^{-2}$	$1,36 \times 10^{-2}$	$1,03 \times 10^{-2}$	$2,41 \times 10^{-2}$

Таблица 25.3-1 дает значения величины (25.3-1) для употребительных скоростей выборки. Исследование таблицы показывает, что из устойчивых методов (кроме метода Симпсона) наименьшую ошибку для низких и умеренных скоростей выборки имеет правило двух третей.

§ 25.4. Дифференциальные уравнения

Имеется существенная разница между вычислением неопределенного интеграла и решением обыкновенного дифференциального уравнения. Чтобы это стало ясно, предположим, что есть дифференциальное уравнение

$$y' = Ay + f(t)$$

и неопределенный интеграл

$$y' = f(t).$$

Для интеграла блок-схема вычислений имеет вид

$$f(t) \rightarrow \boxed{\int} \rightarrow y(t),$$

а для дифференциального уравнения

$$f(t) \rightarrow \begin{array}{c} \boxed{\int} \\ \boxed{A} \end{array} \rightarrow y(t).$$

В последней имеется петля обратной связи.

Это различие делает более подходящим другой подход к интегрированию уравнения, хотя при желании может быть использован метод передаточной функции. Мы будем пользоваться методом исследования величины ошибки на каждом шагу вместо отношения выхода к входу, так как обычно это более существенно для дифференциальных уравнений. В действительности мы лишь даем разные ответы на вопрос «какова точность?», так как разные случаи требуют применения различных критериев точности.

Примем общую формулу (15.2-1) (которая выглядит так же, как и в §§ 25.2 и 25.3, но является другой по смыслу, так как значения y' теперь получаются из значений y через дифференциальное уравнение)

$$y_{n+1} = a_0 y_n + a_1 y_{n-1} + a_2 y_{n-2} + h(b_{-1} y'_{n+1} + b_0 y'_n + b_1 y'_{n-1} + b_2 y'_{n-2}). \quad (25.4-1)$$

Чтобы изучить частотное поведение (25.4-1), подставим функцию

$$y(t) = e^{2\pi i \omega t} = e^{i \omega t} \quad (\omega = 2\pi \sigma)$$

и вычислим ошибку как функцию частоты

$$\begin{aligned} G(\sigma) = G_1(\omega) &= e^{i\omega(t+h)} - a_0 e^{i\omega t} - a_1 e^{i\omega(t-h)} - a_2 e^{i\omega(t-2h)} - \\ &\quad - h i \omega [b_{-1} e^{i\omega(t+h)} + b_0 e^{i\omega t} + b_1 e^{i\omega(t-h)} + b_2 e^{i\omega(t-2h)}], \\ |G_1(\omega)| &= |1 - a_0 e^{-i\omega h} - a_1 e^{-2i\omega h} - a_2 e^{-3i\omega h} - \\ &\quad - i h \omega (b_{-1} + b_0 e^{-i\omega h} + b_1 e^{-2i\omega h} + b_2 e^{-3i\omega h})|. \end{aligned} \quad (25.4-2)$$

Если взять теперь общую функцию

$$y(t) = f(t) = \int_{-\infty}^{\infty} F(\sigma) e^{2\pi i \sigma t} d\sigma,$$

то ошибка будет

$$R[f(t)] = \int_{-\infty}^{\infty} F(\sigma) G(\sigma) e^{2\pi i \sigma t} d\sigma. \quad (25.4-3)$$

Используя теорему свертки, получаем

$$R[f(t)] = \int_{-\infty}^{\infty} f(s) g(t-s) ds = \int_{-\infty}^{\infty} g(t-s) y(s) ds \quad (25.4-4)$$

В этом виде ошибку можно сравнить с ошибкой из гл. 11.

В частности (см. равенство (11.3-8)),

$$R[f(t)] = \int_A^B f^{(m)}(s) G(s) ds,$$

где $G(s)$ была названа функцией влияния и составлена из членов вида

$$(\chi_k - s)^j + \dots$$

Здесь опять выступает важное различие между двумя теориями: многочленный метод приводит к выражению ошибки в виде интеграла от m -й производной функции, умноженной на функцию влияния, тогда как метод интеграла Фурье приводит к использованию самой функции или ее преобразования, а не m -й производной и, конечно, другой функции влияния (см. также (25.2-10)). Однако вид формулы (25.4-4) обманчив; если функцию $g(t-s)$ исследовать тщательнее, то можно увидеть, что это просто замаскированная формула (25.4-1).

§ 25.5. Построение фильтров по методу Чебышева

К первой стадии построения фильтра приводит уже понимание идеи метода. На практике возникает так много различных задач, что нельзя надеяться охватить их все. Мы удовлетворимся поэтому только беглым очерком, проясняющим чебышевский подход к конструкции фильтров. Одна из целей этой книги — привести читателя к тому, чтобы он мог сам применять предложенные принципы и строить формулы. Мы надеемся, что достигли этой цели.

Предположим, что требуется построить такой метод интегрирования дифференциальных уравнений, чтобы ошибка в области частот имела чебышевскую форму, а не форму, имевшуюся ранее — очень хорошую вначале и все ухудшающуюся, когда частота возрастает. Для некоторых целей чебышевский вид ошибки явно предпочтительнее ошибки в виде степенного ряда.

Начнем с преобразования нашей области частот, но так, чтобы она была заключена между -1 и $+1$; после этого можно использовать стандартные чебышевские многочлены. Если принять общую формулу

предыдущего параграфа (равенство (25.4-1)), то придем для ошибки в частоте ω к функции G (равенство (25.4-2))

$$G(\sigma) = e^{i\omega(t+h)} - a_0 e^{i\omega t} - a_1 e^{i\omega(t-h)} - a_2 e^{i\omega(t-2h)} - \\ - i h \omega [b_{-1} e^{i\omega(t+h)} + b_0 e^{i\omega t} + b_1 e^{i\omega(t-h)} + b_2 e^{i\omega(t-2h)}]. \quad (25.5-1)$$

Вспомним теперь гл. 19. Там было замечено, что если нужно иметь ошибку в виде многочлена Чебышева, то следует рассматривать все выражение взятым в виде многочленов Чебышева. Для этого нужно уметь выразить $e^{i\omega z}$ через многочлены Чебышева. В известной книге Ватсона «Теория бесселевых функций» [41] приводится производящая функция

$$e^{z/2 \left(t - \frac{1}{t}\right)} = \sum_{n=-\infty}^{\infty} t^n J_n(z).$$

Положим $t = i e^{i\theta}$ и, используя равенство

$$J_{-n}(z) = (-1)^n J_n(z),$$

или, что то же,

$$(i)^{-n} J_{-n}(z) = (i)^n J_n(z),$$

получим

$$e^{iz \cos \theta} = J_0(z) + 2 \sum_{n=1}^{\infty} (i)^n J_n(z) \cos n \theta.$$

Положим $\cos \theta = \omega$; тогда

$$e^{iz\omega} = J_0(z) + 2 \sum_{n=1}^{\infty} (i)^n J_n(z) T_n(\omega) \quad (-1 \leq \omega \leq 1). \quad (25.5-2)$$

В (25.5-1) вынесем за скобки член $e^{i\omega t}$ и получим

$$|G(\sigma)| = |e^{i\omega h} - a_0 - a_1 e^{-i\omega h} - a_2 e^{-2i\omega h} - i h \omega (b_{-1} e^{i\omega h} + b_0 + \\ + b_1 e^{-i\omega h} + b_2 e^{-2i\omega h})|.$$

Используя (25.5-2) ($z = h, 0, -h, -2h$ по очереди) и соответствующие тождества для $\omega T_n(\omega)$, можно превратить все в разложение по чебышевским многочленам с комплексными коэффициентами (включающими бесселевы функции).

Если приравнять все коэффициенты нулю, то получим одно уравнение для каждого комплексного коэффициента. Можно использовать все эти уравнения, но лучше сохранить свободными один или два параметра для устойчивости и других желаемых свойств. Результатом была бы кривая ошибок, которая не равна нулю при нулевой частоте, что, вероятно, приводит к затруднениям; простое уравнение

$$y' = 0, \quad y(0) = A \neq 0$$

имело бы решение, отличное от постоянной. Чтобы исправить это, можно начать с требования, что нулевая частота удовлетворяется точно, в результате чего

$$1 = a_0 + a_1 + a_2$$

и только затем начать приравнивать коэффициенты нулю. Таким образом, можно добиться уверенности, что в пределах округления нулевая частота вычисляется точно. Заметим, что если сохранить тот же самый интервал частот и изменить скорость выборки за счет изменения h , то меняются все коэффициенты, а не только множитель h , как в случае многочлена.

Итак, мы в основном обрисовали метод Чебышева. Из сказанного видно, как найти желаемую формулу при условии, что мы согласны несколько поработать, чтобы получить ее. И это верно для широкого класса формул — общий метод часто оправдывает себя при условии, что мы сделаем некоторую трудную работу и имеем немного воображения, чтобы преодолеть несколько запутанных мест. Искусство нахождения формул практически сводится к науке; методы, развитые до сих пор, достаточны, чтобы обеспечить больше формул, чем существует места, чтобы перечислить их в книге во много раз большей, чем эта, так как существует много комбинаций идей, которые могут быть использованы при различных обстоятельствах. Есть надежда, что читатель теперь будет чувствовать, что он может с некоторым трудом выводить формулы, соответствующие ситуации, вместо того чтобы подгонять ситуацию под классические формулы; надеемся, что он так и будет поступать.

§ 25.6. Некоторые детали метода Чебышева

Так как методы Чебышева не часто встречаются в литературе, рассмотрим детальнее формулы интегрирования. Пусть формула интегрирования имеет стандартную форму

$$y_{n+1} = a_0 y_n + a_1 y_{n-1} + a_2 y_{n-2} + h (b_{-1} y'_{n+1} + b_0 y'_n + b_1 y'_{n-1} + b_2 y'_{n-2}). \quad (25.6-1)$$

Предположим, как и в методе Фурье, что

$$y(t) = e^{i\omega t};$$

используя уравнение (25.2-2), получим

$$y(t) = J_0(t) + 2 \sum_{n=1}^{\infty} (i)^n J_n(t) T_n(\omega). \quad (25.6-2)$$

Производная равна

$$y'(t) = J'_0(t) + 2 \sum_{n=1}^{\infty} (i)^n J'_n(t) T_n(\omega). \quad (25.6-3)$$

Подставим теперь эти два выражения в (25.6-1), используя соответствующие значения t , а именно $t = h, 0, -h, -2h$, и вынесем за скобку $e^{i\omega h}$:

$$\begin{aligned} J_0(h) = & 2 \sum_{n=1}^{\infty} (i)^n J_n(h) T_n(\omega) = \\ & = a_0 J_0(0) + a_1 [J_0(-h) + 2 \sum_{n=1}^{\infty} (i)^n J_n(-h) T_n(\omega)] + \\ & + a_2 [J_0(-2h) + 2 \sum_{n=1}^{\infty} (i)^n J_n(-2h) T_n(\omega)] + \\ & + b_{-1} h [J'_0(h) + 2 \sum_{n=1}^{\infty} (i)^n J'_n(h) T_n(\omega)] + \\ & + b_0 h [2iJ'_1(0) T_1(\omega)] + b_1 h [J'_0(-h) + 2 \sum_{n=1}^{\infty} (i)^n J'_n(-h) T_n(\omega)] + \\ & + b_2 h [J'_0(-2h) + 2 \sum_{n=1}^{\infty} (i)^n J'_n(-2h) T_n(\omega)]. \end{aligned}$$

Преобразуем это выражение так, чтобы получить разложение по многочленам Чебышева, и приравняем коэффициенты при $T_0(\omega)$:

$$J_0(h) = a_0 + a_1 J_0(h) + a_2 J_0(2h) + h [b_{-1} J'_0(h) - b_1 J'_0(h) - b_2 J'_0(2h)].$$

$$\text{При } \frac{T_1(\omega)}{2i}:$$

$$\begin{aligned} J_1(h) = & -a_1 J_1(h) - a_2 J_1(2h) + h [b_{-1} J'_1(h) + b_0 J'_1(0) + \\ & + b_1 J'_1(h) + b_2 J'_1(2h)], \end{aligned} \quad (25.6-4)$$

$$\text{При } \frac{T_k(\omega)}{2(i)^k}:$$

$$\begin{aligned} J_k(h) = & (-1)^k [a_1 J_k(h) + a_2 J_k(2h)] + h [b_{-1} J'_k(h) - \\ & - (-1)^k b_1 J'_k(h) - (-1)^k b_2 J'_k(2h)]. \end{aligned}$$

Первый (вырожденный) случай, который нужно рассмотреть, эквивалентен правилу трапеций; положим $a_1 = a_2 = b_1 = b_2 = 0$. Потребуем также выполнения условия, чтобы (25.6-1) было точным для $y = 1$; это достигается при

$$a_0 = 1.$$

Теперь возьмем коэффициенты при $T_0(\omega)$ и $T_1(\omega)$:

$$J_0(h) = 1 + h b_{-1} J'_0(h), \quad J_1(h) = h [b_{-1} J'_1(h) + b_0 J'_1(0)].$$

Из них получаем $J'_0(h) = -J_1(h)$,

$$b_{-1} = \frac{1 - J_0(h)}{hJ_1(h)}, \quad b_0 = 2 \frac{J_1(h) - 2hb_{-1}J'_1(h)}{h}. \quad (25.6-5)$$

Для того чтобы понять и проверить эти равенства, исследуем, что происходит при $h \rightarrow 0$. Как известно,

$$J_n(z) = \frac{1}{n!} \left(\frac{z}{2}\right)^n \left[1 - \frac{(z/2)^2}{n+1} + \frac{(z/2)^4}{2!(n+1)(n+2)} - \frac{(z/2)^6}{3!(n+1)(n+2)(n+3)} + \dots \right],$$

так что

$$J_0(h) = 1 - \frac{h^2}{4} + \frac{h^4}{64} - \dots,$$

$$J_1(h) = \frac{h}{2} - \frac{h^3}{16} + \frac{h^5}{384} - \dots,$$

$$J_2(h) = \frac{h^2}{8} - \frac{h^4}{96} + \frac{h^6}{3072} - \dots$$

и

$$b_{-1} = \frac{\frac{h^2}{4} - \frac{h^4}{64} + \dots}{\frac{h^2}{2} - \frac{h^4}{16} + \dots} \approx \frac{1 - \frac{h^2}{16}}{2 \left(1 - \frac{h^2}{8}\right)} \rightarrow \frac{1}{2},$$

$$b_0 = \frac{h - \frac{h^3}{8} + \dots - 2b_{-1}h \left(\frac{1}{2} - \frac{3h^2}{16} + \dots\right)}{h} \rightarrow \frac{1}{2}.$$

Сначала может показаться удивительным, что коэффициенты b_{-1} и b_0 различны, но небольшое размышление показывает, что в этом нет ничего неожиданного. Все дело в том, что мы проводили кривую ошибки через нуль при $\omega = 0$, а потом пытались сделать главный член в разложении ошибки пропорциональным $T_2(\omega)$, которое, очевидно, не имеет нуля при $\omega = 0$. В действительности этот случай вполне тривиален.

Теперь рассмотрим случай, когда используются два старых значения функции и производная, положив лишь $a_2 = b_2 = 0$. В этом случае сохраним для устойчивости один параметр. Таким образом, имеем уравнения:

$$\text{св. чл.:} \quad 1 = a_0 + a_1,$$

$$T_0(\omega): \quad J_0(h) = a_0 + a_1 J_0(h) + h [b_{-1} J'_0(h) - b_1 J'_0(h)],$$

$$T_1(\omega): \quad J_1(h) = -a_1 J_1(h) + h [b_{-1} J'_1(h) + b_0 J'_1(0) + b_1 J'_1(h)],$$

$$T_2(\omega): \quad J_2(h) = a_1 J_2(h) + h [b_{-1} J'_2(h) - b_1 J'_2(h)].$$

Второе уравнение можно записать, исключив a_0 , в виде

$$0 = (1 - J_0)(1 - a_1) + hJ'_0(b_{-1} - b_1),$$

а четвертое уравнение как

$$0 = -J_2(1 - a_1) + hJ'_2(b_{-1} - b_1),$$

откуда видно, что

$$a_1 = 1, \quad b_{-1} = b_1.$$

Мы находимся на границе устойчивости (см. упражнение 13.7-1). Из первого уравнения получаем

$$a_0 = 0.$$

Теперь третье уравнение имеет вид $J_1(0) = \frac{1}{2}$,

$$2J_1 = 2hb_{-1}J'_1 + \frac{hb_0}{2}. \quad (25.6-6)$$

Когда $h \rightarrow 0$, оно переходит в уравнение

$$2b_{-1} + b_0 = 2, \quad (25.6-7)$$

как и должно быть.

Эти результаты напоминают метод Тика (см. § 25.2), который использовал четыре узла на цикл $h = \frac{\pi}{2}$. В своем методе он прибавил условие (25.6-7), означающее, что метод интегрирования точен, когда $u(t)$ есть прямая линия. Используя метод Тика, можно решить уравнения (25.6-6) и (25.6-7), которые дадут

$$b_{-1} = \frac{2J_1(h) - h}{2J'_1(h) - 1} = 0,35785 \quad \left(h = \frac{\pi}{2}\right),$$

что очень хорошо совпадает со значением 0,3584, найденным Тиком. Полученное решение применимо при любой частоте выборки. В этом примере члены более высокого порядка в разложении Чебышева, очевидно, дают очень маленький вклад.

При другом методе можно сделать формулу точной для частот, которые соответствуют нулям $T_3(\omega)$, а именно $\omega = 0$ и $\pm \frac{\sqrt{3}}{2}$; таким образом, кривая ошибки имела бы вид $T_3(\omega)$.

Эти примеры показывают некоторые принципы чебышевского метода. В конкретных случаях, однако, часто имеются особые обстоятельства, которые влияют на выбор формулы, соответствующей данной ситуации. Большое количество таких деталей мешает дальнейшему изложению этого метода в элементарном курсе.

ГЛАВА 26

ЭКСПОНЕНЦИАЛЬНАЯ АППРОКСИМАЦИЯ

§ 26.1. Введение

Из трех классов функций, которые были рассмотрены в § 7.4 и которые инвариантны относительно сдвига независимой переменной — многочленов, синусов и косинусов и экспонент, — мы изучили многочлены в главах с 7 по 20, синусы и косинусы в главах с 21 по 25, и теперь в одной короткой главе поговорим об экспонентах.

Для объяснения такого несоответствия в изложении можно привести несколько доводов. Во-первых, класс многочленов инвариантен также относительно изменения масштаба и, следовательно, в некоторых отношениях более важен. Во-вторых, значительная часть глав о многочленах была посвящена развитию техники, которая может быть использована и для двух других классов. В-третьих, многочлен часто имеет меньше параметров, чем экспоненциальное выражение, что делает эти параметры более удобными для исследования, но менее эффективными при использовании во многих ситуациях.

Верно также и то, что экспоненты могут быть представлены как синусы и косинусы чисто мнимого аргумента, и мы опять находим, что некоторые детали уже исследованы. Еще одна причина состоит в том, что литература и история математики отводят классу экспонент второстепенную роль по сравнению с двумя другими классами. Наконец, переход от частных к общим идеям есть часть плана этой книги, и поэтому мы умышленно опускаем сейчас много деталей.

Задачи экспоненциальной аппроксимации можно грубо разделить на два класса: те задачи, где показатели степеней заданы, и те, где они неизвестны. Рассмотрим сначала случай, когда они заданы.

§ 26.2. О нахождении формул, использующих экспоненты, когда показатели экспонент известны

В гл. 10 был развит достаточно общий метод нахождения формул. Там же он был применен к случаю многочленов; используем этот метод снова.

Предположим, что имеется линейный оператор $L(f)$, например, интегрирования, дифференцирования, интерполяции и т. д. (см. § 10.3), и требуется представить ответ как линейную комбинацию узловых значений с весами (производные исключаются только для удобства)

$$L(f) = w_0 f(x_0) + w_1 f(x_1) + \dots + w_{n-1} f(x_{n-1}). \quad (26.2-1)$$

Здесь имеется n параметров w_0, w_1, \dots, w_{n-1} и, следовательно, можно (мы надеемся) сделать формулу точной для n функций $e^{a_0 x}, e^{a_1 x}, \dots$

..., $e^{a_{n-1}x}$. Часто $a_0 = 0$, так что в число таких функций входит константа.

Уравнения, соответствующие (10.3-1), принимают вид

$$m_k = w_0 e^{a_k x_0} + w_1 e^{a_k x_1} + \dots + w_{n-1} e^{a_k x_{n-1}} \quad (k=0, 1, \dots, n-1), \quad (26.2-2)$$

где, естественно,

$$m_k = L(e^{a_k x}). \quad (26.2-3)$$

Возникает вопрос, можно ли обратить эту систему уравнений; это зависит от значения определителя

$$|e^{a_k x_j}| \quad (k=0, 1, \dots, n-1; \quad j=0, 1, \dots, n-1). \quad (26.2-4)$$

Предположим, что a_k расположены равномерно и, более того, что $a_k = k$. Тогда положим

$$e^{x_j} = u_j \quad (26.2-5)$$

и определитель принимает вид

$$|u_j^k|, \quad (26.2-6)$$

а это — определитель Вандермонда (см. § 8.2), не равный нулю, кроме случая $x_i = x_j$ для некоторого $i \neq j$. Это и не удивительно, так как замена переменной (26.2-5) приводит нас к случаю многочленов.

Рассмотрим теперь общий случай равномерно расположенных значений a_k и положим

$$a_k = a + bk.$$

Тогда (26.2-2) принимает вид

$$m_k = \sum_{j=0}^{n-1} w_j e^{a_k x_j} = \sum_{j=0}^{n-1} (w_j e^{a x_j}) (e^{b x_j})^k.$$

Положив

$$w_j e^{a x_j} = \bar{w}_j, \quad e^{b x_j} = u_j,$$

имеем $m_k = \bar{w}_j u_j^k$, что опять есть случай многочленов. Таким образом, случай равномерно расположенных показателей есть просто замаскированный случай многочленов.

Если решить систему относительно w_k (или \bar{w}_k), то найдем искомую формулу.

Когда показатели известны, но неравномерно расположены, доказательство того, что определитель (26.2-4) отличен от нуля не так просто, и мы не будем здесь дальше рассматривать этот случай. Во всех случаях решение уравнений (26.2-2) дает искомую формулу.

Упражнение 26.2-1. Найдите двумя различными методами формулу

$$\int_0^1 f(x) dx = a_0 f(0) + a_{1/2} f(1/2) + a_1 f(1),$$

которая точна для $f(x) = 1$, e^{-x} и e^{-2x} .

§ 26.3. Неизвестные показатели

Случай интерполяции или представления функции в виде суммы экспонент с неизвестными показателями важен, так как он является основой аналитических замен. Прони дал простой метод нахождения показателей степени для равноотстоящих данных. Пусть

$$f(x) = A_0 e^{a_0 x} + A_1 e^{a_1 x} + \dots + A_{k-1} e^{a_{k-1} x} \quad (26.3-1)$$

для некоторого множества равноотстоящих значений $x = x_j$ ($j = 1, \dots, n$). Не будет ограничением считать, что $x_j = j$.

Прони заметил, что если все члены $e^{a_l x}$ ($l = 0, 1, \dots, k-1$) удовлетворяют некоторому разностному уравнению k -го порядка с постоянными коэффициентами, то характеристические корни этого уравнения равны $\rho = e^{a_l}$. Следовательно, $f(x)$ также удовлетворяет этому разностному уравнению. Пусть это разностное уравнение есть

$$f(j) + c_1 f(j+1) + \dots + c_k f(j+k) = 0 \quad (j = 0, 1, 2, \dots). \quad (26.3-2)$$

Возможны два случая. Если мы имеем ровно столько же уравнений (26.3-2), сколько неизвестных c_m ($m = 1, 2, \dots, k$), то следует рассмотреть «персимметричный» определитель (см. § 10.6)

$$|f(j+n)|.$$

Если он не равен нулю, то можно решить уравнения для c_j . Зная c_j , найдем характеристическое уравнение

$$\rho^k + c_1 \rho^{k-1} + \dots + c_k = 0$$

и из его корней находим a_i . Заметьте, как сильно это напоминает метод решения системы для гауссовой квадратуры (§ 10.6). Теперь, когда известны a_i , можно решить первые k уравнений для A_i . Таким образом, $2k$ равномерно расположенных узлов $f(x)$ определяют $2k$ неизвестных a_i и A_i .

Если имеется больше чем $2k$ узлов, то можно использовать метод наименьших квадратов (главы 17 и 18) и получить нормальные уравнения, соответствующие (26.3-2), из которых найти по очереди a_i и A_i .

Как только есть интерполирующая функция, мы можем произвести аналитическую замену и найти любую формулу, которая нам нужна.

Если мы хотим прийти к ответу, не проходя через интерполирующую функцию, достаточно просто подставить моменты на место зна-

чений функции (26.3-1); естественно, что мы используем столько моментов, сколько неизвестных, а именно $2k$. Опять же общий метод достаточен для нахождения формулы, которую мы хотим иметь, однако следует быть осторожным. Ответ может зависеть от того, в каком месте задачи произведена аппроксимация функции суммой экспонент.

Упражнение 26.3-1. Подобрать функцию $y_1 = A_1 e^{-a_1 x} + A_2 e^{-a_2 x}$ по следующим данным:

0	1	2	3
5,5	5,0	5,1	4,3

§ 26.4. Предупреждения

Хотя в принципе выше было показано, как находить показатели, на практике не всегда все идет так хорошо: иногда случается, что число членов в представлении (26.3-1) не известно, а должно быть найдено. Иллюстрацией этого является радиоактивный распад, где члены соответствуют различным периодам полураспада в цепи распада, который исследуется.

Рассмотрим простой случай попытки различить

$$Ae^{-at} \text{ и } \frac{A}{2} e^{-(a+\varepsilon)t} + \frac{A}{2} e^{-(a-\varepsilon)t} = Ae^{-at} \left[\frac{e^{-\varepsilon t} + e^{\varepsilon t}}{2} \right].$$

Выражение в скобках равно

$$1 + \frac{\varepsilon^2 t^2}{2} + \frac{\varepsilon^4 t^4}{24} + \dots$$

Разность зависит от ε^2 , и только для больших t можно надеяться заметить ее на фоне шумов измерений; но для больших t величина e^{-at} мала! Аналогичная ситуация возникает в преобразовании Лапласа

$$f(t) = \int_0^{\infty} F(\sigma) e^{-\sigma t} d\sigma.$$

Если дано $F(\sigma)$, легко вычислить $f(t)$, но если есть $f(t)$, заданная по точкам, то задача нахождения $F(\sigma)$ более трудна. Ланцош [23] говорит: «... это показывает, что физические наблюдения преобразования Лапласа никогда не могут привести к решению задачи восстановления оригинала с достаточной степенью точности». Одна из трудностей состоит в том, что значения $F(\sigma)$ для больших σ определяют значения $f(t)$ для маленьких t и наоборот. Если $f(t)$ удовлетворяет некоторым

дополнительным условиям или если мы располагаем дополнительной информацией о $F(\sigma)$, то обратное преобразование можно иногда найти.

Читатель не должен отступать, если он встретится с обратным преобразованием Лапласа, так как отдельные результаты иногда можно найти, но удовлетворительные общие методы автору пока неизвестны.

§ 26.5. Экспоненты и многочлены

Когда общее поведение задачи имеет экспоненциальный характер, использование многочлена, возведенного в подходящую степень, часто более удобно, чем сумма экспонент. Метод интегрирования Гаусса — Лагерра и метод, использованный в последней части § 16.5, иллюстрируют это высказывание. Так как общая техника, необходимая для реализации этого предложения, уже изложена во второй части книги, нет нужды обсуждать ее дальше.

§ 26.6. Остаточные члены

Когда формулы найдены, естественно поинтересоваться остаточными членами. В гл. 11 мы начинали с разложения (равенство (11.3-2)) произвольной функции $f(x)$ по функциям $1, x, \dots, x^{m-1}$, для которых формула была сделана точной. Это разложение обобщено на произвольные множества функций; изложение этого вопроса можно найти в превосходной (но трудной для чтения) книге Хаусхолдера «Основы численного анализа» [16].

На практике польза таких остаточных членов сомнительна и они пока редко употребляются. Остаточные члены, соответствующие формулам для функций с ограниченным спектром, по-видимому, еще не исследованы.

ГЛАВА 27

ОСОБЕННОСТИ

§ 27.1 Введение

В § 7.4 был отмечен тот факт, что имеется, по существу, только три класса функций, которые как классы инвариантны относительно сдвига независимой переменной. Эти три класса были рассмотрены во второй части, в главах 21—25 и 26 третьей части, посвященных соответственно многочленам, функциям Фурье ($\sin nx, \cos nx$) и экспонентам.

Значение инвариантности при переносе независимой переменной теряется, если функция имеет особенность, так как расположение осо-

бенности дает естественное начало координат. Кроме того, особенности, характеризующиеся стремлением значений функции к бесконечности для конечных значений x , можно аппроксимировать лишь рациональными функциями. Мы несколько раз были близки к рассмотрению особенностей и указывали в § 16.5, что использование множителя для действия с особенностью обычно лучше, чем использование аддитивного члена. В § 1.9 делалась замена переменного, чтобы устранить особенности.

Использование известного поведения функции вблизи особенности сильно помогает вычислениям. Большая часть этого искусства лежит в области чистой математики и, следовательно, за пределами этой книги. Ограничимся поэтому двумя иллюстративными примерами, после чего сделаем несколько общих замечаний. Однако, как только класс функций определен, процесс нахождения отдельного представителя этого класса следует по тому же пути, который использовался и раньше, а значит, об этом можно больше не говорить.

§ 27.2. Пример интеграла с особенностью в бесконечности

Предположим, что рассматривается интеграл

$$f(x) = \int_0^x e^{x^2} dx.$$

Известно, что его асимптотическое разложение есть

$$f(x) \sim \frac{e^{x^2}}{2x} + \dots,$$

но оно непригодно вблизи $x=0$. Тогда напомним

$$f(x) = \int_0^x e^{x^2} dx = e^{x^2} D(x)$$

и попробуем определить $D(x)$. Продифференцируем это выражение, чтобы получить дифференциальное уравнение:

$$D'(x) + 2xD(x) = 1, \quad D(0) = 0.$$

Теперь нужно проинтегрировать численно простое дифференциальное уравнение, и известно, что $D(x) \rightarrow \frac{1}{2x}$, когда $x \rightarrow \infty$.

Таким образом, функция $D(x)$ хорошо ведет себя на всем интервале $0 \leq x \leq \infty$, и поэтому лучше изучать не заданный интеграл

$$f(x) = \int_0^x e^{x^2} dx = e^{x^2} D(x),$$

а интеграл Дюсона

$$D(x) = e^{-x^2} \int_0^x e^{x^2} dx.$$

Из этого примера видно, как просто иногда можно вычислить интеграл с особенностью. Это справедливо и для случаев, когда функция плохо ведет себя, оставаясь конечной; так, можно применять подобные хитрости, если на комплексной плоскости вблизи действительной оси имеется особенность. В приведенном примере дело свелось к простому введению множителя, уничтожающего особенность, и интегрированию дифференциального уравнения.

§ 27.3. Особенность в линейном дифференциальном уравнении

Иногда бывает, что требуется численное решение линейного дифференциального уравнения, которое имеет особенность*) на отрезке интегрирования. Один из возможных способов проделать это, предложенный впервые профессором Тьюки и снова иллюстрирующий уничтожение особенности с помощью множителя, заключается в следующем. Пусть

$$y'' + P(x)y = 0, \quad y(x_0) = A, \quad y'(x_0) = B, \quad (27.3-1)$$

где $P(x)$ имеет особенность в области интегрирования, например при $x = 0$.

Выберем для сравнения уравнение

$$u'' + Q(x)u = 0, \quad (27.3-2)$$

имеющее известные решения $u_1(x)$ и $u_2(x)$ и особенность $Q(x)$ того же вида и в том же месте, что и $P(x)$. Ниже будет показано, как это можно сделать. Предположим, далее, что

$$y(x) = \alpha(x)u_1(x) + \beta(x)u_2(x), \quad (27.3-3)$$

где $\alpha(x)$ и $\beta(x)$ — функции, которые еще надо определить, но про которые предполагается, что они будут гладкими. Наш метод следует теперь классическому «методу вариации постоянных». Введя две неизвестные функции $\alpha(x)$ и $\beta(x)$, мы можем потребовать выполнения (27.3-1) и еще одного условия, которое выберем так:

$$\alpha'u_1 + \beta'u_2 = 0. \quad (27.3-4)$$

*) То есть какой-либо из коэффициентов уравнения вблизи некоторой точки не ведет себя как степенной ряд.

Таким образом, имеем

$$\begin{aligned}y &= \alpha u_1 + \beta u_2, \\y' &= \alpha u_1' + \beta u_2', \\y'' &= \alpha u_1'' + \alpha' u_1' + \beta u_2'' + \beta' u_2'.\end{aligned}$$

Подставляя эти выражения в (27.3-1), получаем

$$\alpha(u_1'' + Pu_1) + \beta(u_2'' + Pu_2) + \alpha' u_1' + \beta' u_2' = 0. \quad (27.3-5)$$

Далее, используя, что u_1 и u_2 удовлетворяют (27.3-2), находим

$$u_1'' + Pu_1 \equiv (P - Q)u_1, \quad u_2'' + Pu_2 \equiv (P - Q)u_2.$$

Таким образом, из (27.3-5) вытекает

$$\alpha' u_1' + \beta' u_2' = (Q - P)(\alpha u_1 + \beta u_2) = (Q - P)y.$$

Решая это совместно с (27.3-4), приходим к равенствам

$$\alpha'(u_1 u_2' - u_2 u_1') = -u_2(Q - P)y, \quad \beta'(u_1 u_2' - u_2 u_1') = u_1(Q - P)y.$$

Но $u_1 u_2' - u_2 u_1'$ — вронскиан функций u_1 и u_2 , и для этого случая (коэффициент при y' в уравнении равен нулю) он есть константа W_0 *), где

$$W_0 = u_1(x_0) u_2'(x_0) - u_2(x_0) u_1'(x_0).$$

W_0 легко определяется по известным решениям $u_1(x)$ и $u_2(x)$.

Таким образом, необходимо решить уравнения

$$\alpha' = -\frac{u_2(Q - P)y}{W_0}, \quad \beta' = \frac{u_1(Q - P)y}{W_0}, \quad (27.3-6)$$

$$y(x_0) = A, \quad y'(x_0) = B, \quad y = \alpha u_1 + \beta u_2.$$

Требуется выбрать $Q(x)$ так, чтобы множитель $Q - P$ убывал вблизи особенности достаточно быстро, чтобы перекрыть рост произведений $u_2 y$ и $u_1 y$ (или, что то же самое, произведений u_2^2 , $u_1 u_2$ и u_1^2). Если это можно сделать, уравнения, определяющие $\alpha(x)$ и $\beta(x)$ вблизи точки, где была особенность, будут вести себя гораздо лучше.

Покажем теперь, как можно выбрать $u_1(x)$ и $u_2(x)$ для нашего примера. Предположим, что

$$P(x) = \frac{a_{-1}}{x} + a_0 + a_1 x + \dots \quad (a_{-1} \neq 0). \quad (27.3-7)$$

Таким образом, предполагается, что если $y(0) \neq 0$, то $y''(0)$ бесконечно, и, очевидно, невозможно аппроксимировать решение многочленом

*) Это следует из теоремы Ливилля, согласно которой $W(x) = W_0 e^{-\int p_1(x) dx}$, где p_1 — коэффициент при y' . (Прим. ред.)

с любой точностью вблизи точки $x=0$. Попробуем положить

$$u_1 = \frac{x}{1+ax+bx^2} = \frac{x}{D(x)} = xD^{-1}(x). \quad (27.3-8)$$

Тогда

$$u_1'' = [2(D^{-1})' + x(D^{-1})''] \left(\frac{D(x)}{x} u_1 \right) \quad \left(\frac{D(x)}{x} u_1 \equiv 1 \right)$$

или, иначе,

$$u_1'' = \frac{x[2(D')^2 - DD''] - 2DD'}{xD^3} u_1.$$

Но

$$D = 1 + ax + bx^2, \quad D' = a + 2bx, \quad D'' = 2b.$$

Следовательно, для малых x

$$\begin{aligned} u_1'' &= \frac{-2a - 6bx + 2b^2x^2}{xD^3} u_1 = \frac{-2a - 6bx + 2b^2x^2}{x[1 + 2ax + (a^2 + 2b)x^2 + \dots]} u_1 = \\ &= -2 \left[\frac{a}{x} + (3b - 2a^2) + \dots \right] u_1 \end{aligned}$$

Итак,

$$Q = -\frac{2a}{x} + (-6b + 4a^2) + (\dots)x + \dots$$

Сравнивая Q и P , выбираем

$$-2a = a_{-1}, \quad a = \frac{a_{-1}}{2},$$

или

$$-6b + 4a^2 = a_0, \quad b = \frac{a_{-1}^2 - a_0}{6}.$$

При таком выборе a и b в (27.3-8) имеем (используя (27.3-7))

$$Q - P \sim Cx + \dots$$

и $Q - P$ убывает достаточно быстро.

Теперь определим u_2 обычным образом:

$$\begin{aligned} u_2(x) &= u_1(x) \int_{x_0}^x \frac{d\theta}{u_1^2(\theta)} = u_1(x) \int_{x_0}^x \frac{1 + 2a\theta + (a^2 + 2b)\theta^2 + 2ab\theta^3 + b^2\theta^4}{\theta^2} d\theta = \\ &= u_1(x) \left[-\frac{1}{x} + 2a \cdot \ln|x| + (a^2 + 2b)x + abx^2 + \frac{b^2x^3}{3} \right]. \end{aligned}$$

Оставшееся просто. Получившаяся функция $u_1(x)$ не имеет особенности при $x=0$, а $u_2(0) = -1$. Вторая производная функции $u_2(x)$ бесконечна при $x=0$. Произведения $(Q-P)u_1$ и $(Q-P)u_2$ обращаются в нуль при $x=0$, что позволяет хорошо интегрировать урав-

нения (27.3-6) для определения $\alpha(x)$ и $\beta(x)$, так как особенность содержится только в u_2 и отчасти маскируется нулем $Q - P$. Этот пример снова показывает, как полезно выделить главный член разложения в окрестности особенности и затем использовать известную функцию для аппроксимации вблизи особенности.

§ 27.4. Общие замечания

Мы привели два примера того, как обращаться с особенностями в двух частных случаях. В нашем распоряжении нет разработанных правил для общего случая, особенно для нелинейных задач. Многое зависит от математических способностей исполнителя.

На современных машинах с плавающей запятой иногда можно, интегрируя, подойти к особенности так близко, чтобы на оставшемся до особой точки отрезке легко было сделать аналитическую замену. Когда особенность пройдена, можно снова приступить к численному интегрированию. Однако этим следует пользоваться лишь тогда, когда другие методы терпят неудачу, так как в этом случае часто трудно оценивать и контролировать точность.

Когда имеется так называемая «подвижная» особенность, что может случиться в нелинейных задачах, особенно важно быть внимательным к тому, как численные данные используются для подбора функции, аппроксимирующей особенность. Все же такие вещи могут быть и бывали сделаны удачно; это вопрос смелости и тщательности вычислений.

АЛГОРИТМЫ И ЭВРИСТИЧЕСКИЕ МЕТОДЫ

ГЛАВА 28

НАХОЖДЕНИЕ НУЛЕЙ

§ 28.1. Алгоритмы и эвристические методы

До сих пор формулы, которые мы рассматривали, задавали искомую величину в явном виде (исключение составляли только методы прогноза и коррекции). Однако в таких случаях, как нахождение нулей функции или решение системы линейных алгебраических уравнений, величины, которые надо вычислить, заданы неявно. Если в такой ситуации задан определенный процесс для вычисления неизвестных, то говорят, что задан *алгоритм*.

Известно множество частных методов, применяющихся в различных конкретных случаях, но что касается общих принципов, которые можно было бы применять регулярно, дело, надо сказать, находится в примитивном состоянии. Цель этой книги — дать общие методы и идеи, используемые в вычислительной практике. Так как в области алгоритмов их, по-видимому, немного, то нам придется довольствоваться для нескольких выбранных алгоритмов описанием того, что можно, а чего нельзя от них ожидать.

Кроме того, для большинства часто встречающихся задач, требующих применения известных алгоритмов, написаны стандартные программы; и детали методов мало волнуют потребителя, если, конечно, он понимает, что за результаты он получит.

Существует много ситуаций, когда неизвестен алгоритм, который будет давать результат, или известен только метод, состоящий в полном переборе, когда неизвестно, что закончится раньше: задача, машинное время или терпение заказчика.

В такой ситуации приходится прибегать к любому из методов, который выглядит полезным. Положение типично для задач нахождения максимумов или минимумов функций многих переменных. Слишком дорого исследовать, скажем, все локальные минимумы четырнадцатимерного пространства. Мы неизбежно обращаемся к проверкам всяких догадок, удачных идей или любых разумных способов поиска

требуемых величин. Такие процессы называются *эвристическими* и будут рассмотрены в гл. 31. Они лежат на переднем фронте интересных современных исследований в области думающих машин и распознавания образов и по необходимости будут рассмотрены лишь кратко. Таким образом, эта книга идет от явных формул к неявным формулам, для которых известен метод решения (алгоритм), и кончается предложениями, касающимися того, что делать с неявными формулами, для которых практические алгоритмы неизвестны (эвристические методы).

§ 28.2. Метод деления пополам для нахождения корня функции

Если дана непрерывная функция действительного переменного x , которая принимает отрицательное значение при $x = a$ и положительное значение при $x = b$, то известно, что между a и b существует точка, в которой функция обращается в нуль. Если разделить интервал пополам и определить, положительно или отрицательно функция в точке деления, то тем самым найдется подынтервал, в котором функция меняет знак (или нуль функции). В принципе, повторным применением этого метода (деление пополам) можно сколь угодно близко подойти к корню. Так как каждый шаг делит интервал, в котором лежит корень, пополам, то 10 шагов уменьшат интервал приблизительно в 1000 раз, 20 шагов — в 1 000 000 раз и т. д. Этот метод, который предполагает только непрерывность и возможность вычисления функции в любой точке, вполне эффективен. Следует сказать: «в принципе», потому что существуют погрешности вычислений.

Давайте посмотрим, как эти погрешности отражаются на процессе нахождения нуля. Предположим, что при вычислении значения функции получается не тот знак из-за погрешностей округления. Результатом явится то, что на следующем шаге будет рассмотрена не та половина интервала. Однако, если нет других ошибок в определении знака функции, всегда будет получаться, что корень заключен в подынтервале, один конец которого находится в точке, где был неверно определен знак. И это вполне приемлемый результат: местонахождение корня определяется в пределах наших возможностей вычислять саму функцию. Небольшая погрешность не привела к слишком большой ошибке.

Каждый предлагаемый метод должен обязательно исследоваться на устойчивость по отношению к ошибкам округления. Многие из наиболее быстрых методов приводят к большим ошибкам, когда они работают при помехах. Следует помнить, что случайно можно оказаться вблизи нуля и, следовательно, оказаться подверженными ошибкам из-за помех на любом этапе вычислений. Отметим еще раз, что

метод деления пополам требует только умения вычислять значения функции и не предполагает составления программы для вычисления производной.

Упражнение 28.2-1. Составить блок-схему программы метода деления пополам.

§ 28.3. Линейная интерполяция

Вероятно, читателю сразу придет в голову, что метод деления пополам можно значительно улучшить, если использовать для следующего вычисления не середину отрезка, а то значение x , в котором дает нуль линейная интерполяция между двумя известными значениями противоположного знака.

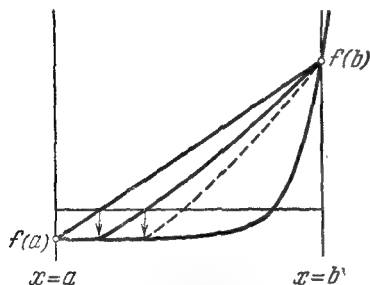


Рис. 28.3-1. Пример медленной сходимости.

Рис. 28.3-1 показывает функцию, для которой метод линейной интерполяции будет наверняка медленнее, чем метод деления пополам.

Конечно, можно возразить, что для разумного множества функций интерполяционный процесс будет в среднем требовать нахождения меньшего числа значений функции. Однако основное достоинство метода деления пополам состоит в том, что здесь извест-

но заранее, сколько шагов потребуется, чтобы получить заданную точность в положении корня; другие методы в среднем могут требовать меньшего числа шагов, однако никогда нельзя сказать, что это будет так в каждом конкретном случае. Действительно, как видно из рис. 28.3-1, можно построить функцию, для которой необходимое число шагов будет больше любого наперед заданного. Для этого достаточно увеличить значение функции в точке $x=b$. Другой метод линейной интерполяции, в которой используются два последних вычисленных значения, если кривая достаточно плохая, может привести при последовательных шагах к удалению от корня.

Упражнение 28.3-1. Составить блок-схему для метода линейной интерполяции.

§ 28.4. Параболическая интерполяция

Если линейная интерполяция лучше, чем метод деления пополам, то, возможно, еще лучше метод построения параболы по трем точкам. Во всяком случае, его следует рассмотреть.

Как и в методе деления пополам, произвольно разделим интервал для нахождения третьей точки. (Может быть, предпочтительнее другие

стратегии, но мы не станем их обсуждать.) Построим теперь параболу, используя разделенные разности и, например, формулу Ньютона (см. § 8.4). Нуль этой параболы, который попадает в интервал, дает нам четвертую точку.

Здесь есть по крайней мере две возможности: можно взять две точки, наиболее близкие к только что полученному значению корня, или, из осторожности, взять две близкие точки, в которых функция имеет разные знаки. Автор предпочитает последнее на том основании, что он любит иметь нуль, который он ищет, заключенным внутри известного интервала и чтобы была уверенность, что парабола имеет действительный нуль в этом интервале. На первом шаге обе возможности дают один и тот же результат, но при последующих шагах эти способы могут дать различные результаты.

Предоставляем читателю проверить влияние помех квадратичной интерполяции при приближении к корню.

Последние два метода иллюстрируют несколько неопределенный общий принцип: *чем тоньше метод и чем лучше он кажется, тем хуже он может повести себя в случае осложнений с функцией*. Он может и в самом деле оказаться хуже простых методов и, вероятно, будет более подвержен помехам.

Некоторое общее представление о том, что нами сделано, можно получить, вернувшись к четырем основным вопросам, поставленным в § 7.1:

1. Каковы узлы?
2. Каков класс функций?
3. Каков критерий согласия?
4. Какова ошибка?

Узлы выбирались здесь по ходу дела, для аналитической замены использовались интерполяционные многочлены, а мерой успеха служила ошибка по x ; для перехода к ошибке по y потребовались бы совсем небольшие изменения.

Очевидно, для интерполяции можно было использовать другие классы функций, если они более отвечают природе изучаемых функций. В нужных случаях это действительно следует делать.

Упражнения

28.4-1. Начертите блок-схему квадратичной интерполяции, сохраняющей перемену знака.

28.4-2. Исследуйте влияние помех на метод квадратичной интерполяции.

28.4-3. Что бы вы назвали ошибкой, контролирующей либо x , либо y ?

§ 28.5. Некоторые общие замечания

Мы использовали линейную и квадратичную интерполяцию для аппроксимации нулей функции. Очевидно, можно идти дальше и строить многочлены более высоких степеней. Выбор стратегии зависит от того, насколько трудно вычислять функцию. Если это очень

трудно и требуются, может быть, часы машинного времени на каждую точку, то, вероятно, следовало бы обратиться к многочленам более высокой степени, повышая степень с каждой следующей точкой. С другой стороны, если функция считается легко, то лучше предпочесть какой-нибудь простой испытанный метод, как, например, деление пополам.

Обычно, когда для нахождения нулей функции $y = y(x)$ используют многочлены, исходные данные аппроксимируют многочленом от x и затем применяют обратную интерполяцию для нахождения нуля. Часто столь же разумно предположить, что x есть многочлен от y , и применять прямую интерполяцию. Безусловно, это намного легче.

Мы рассмотрели также (§ 7.1) метод Ньютона и в упражнении 7.1-1 указали на более общий метод, который использует вторую производную. А. М. Островский в своей великолепной книге [34] предполагает, что вычисление функции должно быть принято за единицу работы, а вычисление каждой производной должно рассматриваться как другая единица работы.

Отсюда он оценивает линейную интерполяцию по двум последним величинам в 1,618 единицы, имея в виду, что каждый шаг увеличивает число верных знаков в 1,618 раза, в то время как метод Ньютона оценивается в 1,414 единицы на единицу работы. Отсюда можно сделать вывод, что метод линейной интерполяции более эффективен. Но, как неоднократно подчеркивалось, коль скоро функция вычислена, после этого большей частью есть почти все необходимое для вычисления производной, и чтобы получить ее, требуется еще очень немного работы; таким образом, если принимается во внимание только машинное время, то вполне вероятно, что метод Ньютона сравним с линейной интерполяцией или даже лучше ее. С другой стороны, метод Ньютона требует аналитического нахождения производной (а при этом человек нередко ошибается) и дополнительного программирования.

Следует осветить еще один вопрос. Обычно оценивают различные методы по их конечной скорости сходимости к истинному решению (в математическом смысле) и пренебрегают вопросом о том, сколько шагов придется сделать, прежде чем такая скорость будет достигнута. И снова опыт показывает, что некоторые из наиболее быстрых методов требуют очень большого времени на подготовку. Примеры, приводимые в учебниках, обычно имеют от 10 до 20 десятичных знаков. Однако на практике в большинстве случаев требуется от трех до пяти десятичных знаков, так что такие примеры часто вводят в заблуждение.

Наконец, молчаливо предполагалось, что ноль с самого начала изолирован и читатель хотел бы знать, как сделать этот первый существенный шаг. К сожалению, автор не может сказать на этот счет

ничего действительно полезного. Часто задача сама по себе при внимательном изучении дает некоторую информацию и существует несколько математических теорем на этот предмет. Но вообще отделение корней есть искусство.

§ 28.6. Метод Берстоу для нахождения комплексных корней многочлена

Задача нахождения корней многочленов возникает достаточно часто для того, чтобы оправдать ее тщательное изучение и разработку специальных методов для ее решения. Мы будем рассматривать только многочлены с действительными коэффициентами, так как обычно с ними и сталкиваются на практике. Фундаментальным свойством действительных многочленов является то, что они могут быть разложены на действительные линейные и действительные квадратичные множители. Предположим, что действительные линейные множители уже удалены. Комплексными корнями следует заниматься после того, как все действительные найдены.

Различным известным методам нахождения действительных линейных и квадратичных множителей можно посвятить целую книгу. Тот факт, что существует так много методов, показывает, что не существует ни одного вполне удовлетворительного; каждый математик имеет свой собственный излюбленный метод или методы. По-видимому, метод Берстоу употребляется вообще чаще любого другого. Он не непогрешим; временами он медленно сходится или даже не сходится вовсе. Но в среднем он кажется лучше любого другого метода *). Вероятно, библиотека стандартных программ каждого вычислительного центра всегда имеет некоторую уже готовую программу, и вы скорее будете использовать ее, чем писать свою собственную. Но все же стоит обсудить один метод, который сильно похож на многие другие.

Пусть дан многочлен

$$P(x) = a_0 + a_1x + a_2x^2 + \dots + a_nx^n. \quad (28.6-1)$$

Предположим, что у нас есть какие-либо догадки о квадратичном множителе

$$x^2 + px + q, \quad (28.6-2)$$

являющемся делителем этого многочлена (вначале можно полагать $p = q = 0$, что упростит первый шаг). Используя деление многочле-

*) Употребителен также метод Мюллера, некоторые предпочитают метод Лина.

нов (см. § 1.6), разделим многочлен на множитель, получая частное и остаток, т. е.

$$a_n x^n + a_{n-1} x^{n-1} + \dots + a_0 = (x_2 + px + q)(b_n x^{n-2} + b_{n-1} x^{n-3} + \dots + b_2) + b_1 x + b_0. \quad (28.6-3)$$

Смысл индексов у b станет ясным в дальнейшем; это облегчает систему обозначений. По схеме деления многочленов имеем

$$\begin{array}{r} \begin{array}{c} 1 \quad p \quad q \\ \hline \end{array} \begin{array}{cccccccc} a_n & a_{n-1} & a_{n-2} & a_{n-3} & \dots & a_2 & a_1 & a_0 \\ - & & qb_n & qb_{n-1} & \dots & qb_1 & qb_2 & qb_3 \\ \hline - & pb_n & qb_{n-1} & & & pb_3 & pb_2 & - \\ \hline b_n & b_{n-1} & b_{n-2} & & \dots & b_2 & \underline{b_1} & b_0 \end{array} \end{array}$$

где остаток равен

$$b_1 x + b_0.$$

Коэффициенты связаны следующими алгебраическими уравнениями:

$$\left. \begin{array}{l} b_n = a_n \\ b_{n-1} = a_{n-1} - pb_n \\ b_{n-2} = a_{n-2} - pb_{n-1} - qb_n \\ \dots \dots \dots \\ b_{n-k} = a_{n-k} - pb_{n-k+1} - qb_{n-k+2} \quad (k=2, 3, \dots, n-1), \\ \dots \dots \dots \\ b_0 = a_0 - qb_2. \end{array} \right\} \quad (28.6-4)$$

Искомый квадратичный множитель получается тогда и только тогда, когда остаток будет тождественно равен нулю, т. е.

$$b_1 = b_0 = 0.$$

Рассмотрим эти коэффициенты как функции p и q ,

$$b_1 = b_1(p, q), \quad b_0 = b_0(p, q).$$

Воспользуемся теперь двумерным аналогом метода Ньютона (см. § 7.1) и разложим b_1 , b_0 в ряд Тейлора вблизи выбранной точки p, q . Обозначив искомое решение через p^* , q^* , получим

$$\left. \begin{array}{l} b_1(p^*, q^*) = 0 = b_1(p, q) + \frac{\partial b_1}{\partial p} \Delta p + \frac{\partial b_1}{\partial q} \Delta q + \dots, \\ b_0(p^*, q^*) = 0 = b_0(p, q) + \frac{\partial b_0}{\partial p} \Delta p + \frac{\partial b_0}{\partial q} \Delta q + \dots, \end{array} \right\} \quad (28.6-5)$$

где

$$\Delta p = p^* - p, \quad \Delta q = q^* - q \quad (28.6-6)$$

суть ошибки, которые должны быть добавлены (приближенно) для получения следующих p и q . Отбрасывая в (28.6-5) все члены, кроме линейных, получаем пару линейных уравнений для приращений, которые должны быть сделаны по p и q .

Задача сводится к нахождению частных производных, которые являются коэффициентами при неизвестных Δp и Δq . Можно было бы дать небольшое приращение p и определить приращения b_1 и b_0 и то же самое проделать для приращения q . Проще, однако, найти их аналитически. Дифференцируем уравнения (28.6-4) по p :

$$\begin{aligned}\frac{\partial b_n}{\partial p} &= 0, \\ \frac{\partial b_{n-1}}{\partial p} &= -b_n - p \frac{\partial b_n}{\partial p}, \\ \frac{\partial b_{n-2}}{\partial p} &= -b_{n-1} - p \frac{\partial b_{n-1}}{\partial p} - q \frac{\partial b_n}{\partial p}, \\ &\dots \dots \dots \\ \frac{\partial b_{n-k}}{\partial p} &= -b_{n-k+1} - p \frac{\partial b_{n-k+1}}{\partial p} - q \frac{\partial b_{n-k+2}}{\partial p}, \\ &\dots \dots \dots \\ \frac{\partial b_0}{\partial p} &= \dots \dots \dots - q \frac{\partial b_2}{\partial p}.\end{aligned}$$

Если теперь обозначить

$$\frac{\partial b_k}{\partial p} = -c_k^* \quad (28.6-7)$$

то получим

$$\left. \begin{aligned}c_n^* &= 0, \\ c_{n-1}^* &= b_n - p c_n^*, \\ c_{n-2}^* &= b_{n-1} - p c_{n-1}^* - q c_n^*, \\ &\dots \dots \dots \\ c_{n-k}^* &= b_{n-k+1} - p c_{n-k+1}^* - q c_{n-k+2}^*, \\ &\dots \dots \dots \\ c_0^* &= \dots \dots \dots - q c_2^*.\end{aligned} \right\} \quad (28.6-8)$$

Эти уравнения имеют ту же форму, что и уравнения (28.6-4), кроме $c_n^* = 0$, и могут быть получены отсюда, если в уравнениях (28.6-4) подставить c_{n-k}^* взамен b_{n-k+1} и b_k взамен a_{k-1} . Только последнее уравнение придется при этом слегка подправить.

Эти замечания наводят на мысль повторить процесс деления многочленов для b (вместо a), используя тот же самый квадратичный

Мы сосредоточим свое внимание на двух широко используемых методах, которые типичны для двух больших классов. Выбор метода, который наиболее соответствует данной машине, есть сложный вопрос.

Говорят, что система линейных уравнений является плохо обусловленной, если, грубо говоря, уравнения почти линейно зависимы. Много усилий ушло на изучение того, как решать плохо обусловленные системы. Однако можно поставить вопрос: требуется ли решать такие системы в практических ситуациях? В какой физической ситуации окажутся полезными ответы, когда они так ощутимо зависят от коэффициентов системы? Обычно справедливо следующее: вместо ответа ищут минимальную систему почти линейно независимых уравнений. В свете этой информации задача может быть лучше понята и обычно переформулирована снова более удовлетворительным образом. Вполне вероятно, что плохо обусловленные системы уравнений, если исключить ошибки округления и измерения, являются действительно линейно зависимыми и, следовательно, не отражают физической ситуации.

§ 29.2. Метод исключения Гаусса

Из *прямых* методов решения систем линейных уравнений, в противоположность *итерационным* методам, наиболее широко используется метод Гаусса. Он состоит в том, что решение системы линейных уравнений осуществляется исключением переменных по очереди. Существует множество различных вариантов метода. Ниже будет рассмотрен один из них. Так как метод является основным, рассмотрим сначала очень простой пример. Пусть дана система:

$$\begin{aligned}x + 2y + 3z &= 10, \\x + 3y - 2z &= 7, \\2x - y + z &= 5.\end{aligned}$$

Вначале исключаем x вычитанием первого уравнения, умноженного на подходящие числа из второго и третьего уравнений. Это дает

$$\begin{aligned}x + 2y + 3z &= 10, \\y - 5z &= -3, \\-5y - 5z &= -15.\end{aligned}$$

Затем исключаем неизвестное y вычитанием второго уравнения, умноженного на -5 , из третьего:

$$\begin{aligned}x + 2y + 3z &= 10, \\y - 5z &= -3, \\-30z &= -30.\end{aligned}$$

Здесь уместны некоторые замечания. В принципе, если система достаточно велика, то наше первоначальное выравнивание может полностью исчезнуть и время от времени может понадобиться повторное выравнивание. Но мы не будем рассматривать такие большие системы. Если системы имеют 50 или более уравнений, то, прежде чем тратить машинное время, лучше проконсультироваться у специалистов. Дело в том, что когда система превосходит по размеру 10 или 20 уравнений, часто начинают появляться новые эффекты и опасности, в которые мы здесь вникать не будем.

Когда выбирается уравнение, которое будет использовано для исключения следующей переменной, можно сначала разделить все коэффициенты этого уравнения на старший коэффициент, так чтобы первый коэффициент стал равен единице. Тогда при обратном ходе нет необходимости ни на что делить. Если деление требует много машинного времени, то можно подсчитать обратную величину к выбранному главному элементу и получить множители (все ≤ 1) для исключения неизвестных из этого уравнения.

Можно было бы подумать, что следует, вместо того чтобы брать x_k подряд, искать очередное x_k , выбирая наибольший коэффициент из всех оставшихся; но эта дополнительная тонкость как будто не увеличивает точности.

§ 29.3. Варианты метода Гаусса

Существует много вариантов метода Гаусса. Однако наиболее значителен метод, называемый методом полного исключения или методом Гаусса — Жордана. В этом методе, когда исключают x_k , то исключают его из всех уравнений, включая и те, которые разрешены для предыдущих переменных. При этом способе нет необходимости в обратном ходе. Иногда утверждают, что этот метод имеет преимущества в точности.

В случае симметричных коэффициентов, если главные элементы берутся на главной диагонали, работа может быть сокращена почти вдвое, но может возникнуть некоторая потеря точности из-за плохого выбора главных элементов. Если матрица положительно определенная, то потери точности малы.

Многие варианты метода обычно являются незначительными изменениями, сделанными для удовлетворения характеристикам данной машины (хотя они могут оказать заметное влияние на полученные результаты); поэтому они не подходят для обсуждения в начальном курсе. Хорошим пособием является книга «Современные методы вычисления» [29] *); в ней в свою очередь есть обширная библиография по многим аспектам вычислений, включая системы уравнений.

*) См. также [16].

§ 29.4. Метод Гаусса — Зайделя

В противоположность прямым методам решения, таким как метод Гаусса, существуют итерационные методы, такие как метод Гаусса — Зайделя. Переставим уравнения в примере § 29.2 так, чтобы больший коэффициент первого уравнения был бы при x , больший коэффициент второго уравнения — при y и третьего — при z :

$$\begin{aligned} 2x - y + z &= 5, \\ x + 3y - 2z &= 7, \\ x + 2y + 3z &= 10. \end{aligned}$$

Начинаем с приближенного решения, скажем, $x = y = z = 0$. Используем теперь первое уравнение для нахождения нового значения

$$x = \frac{5 + y - z}{2} = \frac{5}{2}.$$

Беря $x = \frac{5}{2}$, $z = 0$, решаем второе уравнение относительно y

$$y = \frac{7 - x + 2z}{3} = \frac{3}{2}.$$

Наконец, используя эти вычисленные значения, решаем третье уравнение относительно z

$$z = \frac{10 - x - 2y}{3} = \frac{3}{2}.$$

Эти три величины дают новое приближение и можно повторить цикл:

$$x = \frac{5}{2}, \quad y = \frac{5}{2}, \quad z = \frac{5}{6}.$$

Следующее повторение дает

$$x = \frac{10}{3} = 3,33 \rightarrow 3, \quad y = \frac{16}{9} = 1,780 \rightarrow 2, \quad z = \frac{28}{27} = 1,04 \rightarrow 1.$$

Заметим, что сходимость медленная. Если возникает ошибка, то она может повлиять на число шагов, но не влияет (в принципе) на конечный ответ. Обычно продолжают итерации до тех пор, пока изменения не станут достаточно малы; что это может означать по отношению к ответу — уже другой вопрос.

Общий случай во многом такой же. Уравнения располагаются так, чтобы большие коэффициенты были на главной диагонали. Если эти члены главной диагонали достаточно сильно превосходят другие коэффициенты уравнения, то сходимость гарантирована, в противном случае — нет. Вот одно достаточное условие:

$$|a_{ii}| > |a_{i1}| + |a_{i2}| + \dots + |a_{ii-1}| + |a_{ii+1}| + \dots + |a_{in}|$$

для всех i (предполагается, что система не распадается на две независимые системы уравнений).

Если в системе имеется много нулевых коэффициентов a_{ij} , то этот метод предпочтительнее метода исключения. Метод Гаусса — Зайделя использует существование нулей, в то время как метод исключения — нет.

Существует множество различных вариантов итерационного метода. При медленной сходимости мы стараемся угадать, используя различные технические приемы, где она начинается, и попасть туда за один шаг. Так называемые *релаксационные* методы являются методами такого типа. За деталями снова отсылаем читателя к [29].

§ 29.5. Повышенная точность

Если имеется некоторый метод решения системы уравнений и нужно повысить точность, то можно воспользоваться следующим приемом. Подставим в уравнение найденные величины x и вычислим разности, используя двойную точность. Затем снова решим систему уравнений, подставляя полученные разности в правые части уравнений. Сумма нового решения со старым дает более точное решение. Легко видеть, почему это так. Первое множество значений $x^{(1)}$ удовлетворяет уравнениям с правыми частями, равными b_i минус разность r_i в i -м уравнении, тогда как второе решение $x^{(2)}$ удовлетворяет приблизительно уравнениям с разностями в правых частях. Следовательно, $x^{(1)} + x^{(2)}$ удовлетворяет тем же самым уравнениям, но с суммой двух правых частей:

$$(b_i - r_i) + r_i = b_i$$

§ 29.6. Общие замечания

Очевидно, здесь не нужно было брать никакого класса функций для аналитической замены. Единственный уместный вопрос из схемы гл. 7 — четвертый: *какова точность?* Хотим ли мы, чтобы x_i были точными? Хотим ли мы, чтобы разности были малы? Хотим ли мы, чтобы какая-то система уравнений, для которой это подсчитанное решение является точным решением, была близка к начальной системе? Последнее условие звучит несколько странно, так как коэффициентов $n(n+1)$, а ответов только n ; n их отклонений от истинного решения могут быть распределены среди $n(n+1)$ коэффициентов многими и многими способами, и отклонения правых частей есть всего лишь частный случай. Таким образом, вопрос в этом случае существенно отличается от того же самого вопроса для нулей многочлена, где мы должны были распределить n ошибок в нулях среди n коэффициентов. Представляется, что на вопрос о точности в данном случае ответить труднее, чем в большинстве других случаев. Часто лучше всего свидетельствуют о точности маленькие разности по отношению к данным правым частям.

ГЛАВА 30

ОБРАЩЕНИЕ МАТРИЦ И СОБСТВЕННЫЕ ЗНАЧЕНИЯ

§ 30.1. Введение

Тщательное изучение матриц ведется уже давно, и про них известно очень много. Они встречаются и во многих различных физических задачах. В результате на эту тему имеется огромная литература и многочисленные методы для различных задач. Рассмотрим лишь несколько примеров и отошлем читателя за дальнейшими сведениями к следующим книгам: Ральстон и Уилф [37], «Современные вычислительные методы» [29] и Хаусхолдер [16], главным образом к библиографии в двух последних.

Мы предполагаем, что читатель знаком с элементарной теорией матриц.

§ 30.2. Обращение матрицы методом исключения по Гауссу

Часто бывает необходимо обратить квадратную матрицу

$$A = (a_{ij}).$$

Представим себе прямоугольную матрицу

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} & 1 & 0 & \dots & 0 \\ a_{21} & a_{22} & \dots & a_{2n} & 0 & 1 & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \dots & a_{nn} & 0 & 0 & \dots & 1 \end{pmatrix},$$

которая получается, если к A приписать единичную матрицу I .

Применим теперь метод исключения Гаусса, не заботясь о правых столбцах. Когда гауссово исключение закончится, получим

$$\begin{pmatrix} 1 & 0 & \dots & 0 & b_{11} & b_{12} & \dots & b_{1n} \\ 0 & 1 & \dots & 0 & b_{21} & b_{22} & \dots & b_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & 1 & b_{n1} & b_{n2} & \dots & b_{nn} \end{pmatrix}.$$

Утверждается, что матрица B

$$B = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ b_{n1} & b_{n2} & \dots & b_{nn} \end{pmatrix}$$

есть матрица, обратная к A . Чтобы убедиться в этом, заметим, что каждый шаг процесса исключения эквивалентен умножению слева на некоторую матрицу. Произведение всех этих левых матриц есть, очевидно, A^{-1} , потому что оно приводит A к единичной матрице I . Но, будучи применено к n правым столбцам, это произведение делает из единичной матрицы матрицу B :

$$A^{-1} \cdot I = B.$$

Следовательно, B и есть обратная к A матрица.

§ 30.3. Задача нахождения собственных значений

Многие задачи приводят к системе уравнений:

$$a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = \lambda x_1,$$

$$a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = \lambda x_2,$$

$$\dots \dots \dots$$

$$a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n = \lambda x_n,$$

или

$$Ax = \lambda x,$$

где λ неизвестно. Эти уравнения совместны тогда и только тогда, когда

$$\Delta = \begin{vmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} - \lambda \end{vmatrix} = 0.$$

Определитель есть многочлен от λ степени n , и, в принципе, если вычислить этот определитель для $n+1$ значения λ , можно использовать эти значения для нахождения многочлена

$$\Delta(\lambda) = 0.$$

Таким путем находятся, в конце концов, и корни многочлена $\Delta(\lambda)$. Впрочем, ошибки округления и большое число необходимых вычислений обычно заставляют отказаться от этого метода в пользу других.

Иногда нужно только наибольшее по модулю значение λ . Предположим, что λ действительное. В таком случае можно действовать следующим образом. Возьмем произвольный вектор

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}.$$

Известно, что если собственные значения различны, то соответствующие собственные векторы (решения системы уравнений для соответствующих собственных значений) образуют полную систему, т. е. любой вектор y можно представить в виде

$$y = c_1 y_1 + c_2 y_2 + \dots + c_n y_n,$$

где y_i — собственные векторы, а c_i — константы.

Умножим обе части равенства на A

$$Ay = \sum_{i=1}^n c_i Ay_i = \sum_{i=1}^n c_i \lambda_i y_i,$$

и вообще

$$A^k y = \sum_{i=1}^n c_i \lambda_i^k y_i.$$

Теперь, если $|\lambda_1| > |\lambda_l|$ для $l = 2, 3, \dots, n$, то при $k \rightarrow \infty$

$$A^k y \approx c_1 \lambda_1^k y_1.$$

Рассмотрим отношение

$$\frac{A^{k+1}y}{A^k y} \approx \frac{c_1 \lambda_1^{k+1} y_1}{c_1 \lambda_1^k y_1} = \lambda_1.$$

Можно ожидать, что в пределе каждая компонента нашего вектора будет умножаться на одно и то же число λ_1 .

На практике рекомендуется нормировать получающийся вектор на каждом шаге так, чтобы наибольшая компонента его была равна 1. На последнем шаге нормирующий множитель даст величину наибольшего по модулю собственного значения. При этом заодно получается собственный вектор.

Этот процесс, очевидно, можно ускорить, если предварительно образовать какую-либо степень A . Вычислим сначала

$$A^2, A^4, A^8, \dots, A^{2^k}.$$

Тогда, применяя

$$A^{2^k} y = \sum_{i=1}^n c_i \lambda_i^{2^k} y_i,$$

получим более быструю сходимость к собственному вектору. Корень степени 2^k нормирующего множителя на последнем шаге даст λ_1 .

Вопрос о том, что будет, если случайно $c_1 = 0$, несерьезен. Если даже это и было верно вначале, то случайная ошибка сделает это слагаемое ненулевым позже и, в конце концов, оно станет домини-

рующим. Это показывает, что удачная догадка о структуре собственного вектора с наибольшим собственным значением ускоряет процесс лишь на первых шагах.

Мысль, что, зная уже наибольшее собственное значение и вектор, мы можем аккуратно вычитать его на каждом шаге и тем самым дать возможность проявиться второму по модулю собственному значению, приходит в голову почти каждому. Это действительно можно сделать, но отнюдь не в точности так, как хотелось бы. На самом деле можно найти несколько старших собственных значений, пока процесс не превратится в шум, так что каждое следующее собственное значение будет определено все с меньшей точностью.

§ 30.4. Наименьшие собственные значения

Чтобы тем же методом найти наименьшее (в алгебраическом смысле) собственное значение, достаточно следующего простого наблюдения. Пусть y_i — собственный вектор, т. е.

$$Ay_i = \lambda y_i.$$

Тогда $(A - pI)y_i = (\lambda - p)y_i$.

Подходящим выбором p можно наименьшее собственное значение сделать наибольшим (по модулю). Пусть собственные значения примерно такие:

$$1, 2, \dots, 10.$$

Выберем $p \approx 10$; тогда $\lambda - p$ будут

$$-9, -8, \dots, 0.$$

Если уже известна примерная величина наибольшего собственного значения, то можно взять p равным этой величине с обратным знаком и самое маленькое собственное значение станет самым большим (по модулю).

§ 30.5. Несколько замечаний

Теория нахождения сразу всех собственных значений сложна и ее нельзя как следует изложить в элементарном курсе. В настоящее время для симметричных матриц, по-видимому, лучшим из прямых методов является модификация Хаусхолдера метода Гивенса.

Хотя существуют прямые методы для несимметричных матриц, с ними надо быть осторожным из-за внутренней неустойчивости задачи, и если матрица велика, то лучше проконсультироваться со специалистом, прежде чем растрачивать машинное время. Мы отсылаем читателя к трем книгам, упомянутым в § 30.1.

ГЛАВА 31

НЕКОТОРЫЕ ПРИМЕРЫ МОДЕЛИРОВАНИЯ

§ 31.1. Введение

Идея моделирования интуитивно ясна, но, по-видимому, не имеет никакого удовлетворительного определения. Поэтому придется воспользоваться примерами.

Моделирование как процессов, так и отдельных ситуаций обычно тесно связано с оптимизацией. В то время как собственно моделирующая часть обычно прозрачна, для осуществления оптимизации часто не известен никакой практически пригодный алгоритм. Цель следующих примеров — изложить, как действовать в ситуации, где неизвестен алгоритм решения задачи. Занятие такими задачами непосредственно приводит в популярную область «думающих машин», где большое внимание привлекают игра в шахматы, доказательство теорем и т. п. Мы, однако, не будем заходить слишком далеко в этой важной и активно развивающейся области науки.

Задачи моделирования занимают значительную часть времени на многих цифровых машинах, и поэтому их нельзя игнорировать только из-за отсутствия общих методов и глубоких результатов. Довольно очевидно, что каждый может сказать: «Давайте точно промоделируем этот процесс», и гораздо труднее специалисту показать, что во многих случаях точное моделирование — это совсем не то, что требуется, а просто большая потеря машинного времени (и, следовательно, денег). Ввиду отсутствия общих правил относительно того, что, когда и как моделировать (и моделировать ли вообще), мы приведем приблизительные соображения общего характера. По мере того как будут рассматриваться примеры, мало отличающиеся от реальных, станет ясно, что эти соображения не приводят к содержательной и полной теории.

Существует грубое, но полезное деление моделирования на *дискретное* и *непрерывное*, соответственно ситуациям, в которых интересующие нас величины в основном дискретны или непрерывны. К сожалению, между ними не всегда есть резкая граница. Практически можно принять разделение, основанное на том, возникают или нет серьезные трудности из-за ошибок замены непрерывной величины дискретной.

Займемся сначала задачами, где таких трудностей нет. Позже будут кратко рассмотрены некоторые детали моделирования, включающего учет таких ошибок, но большая часть относящегося сюда материала уже изложена, главным образом в гл. 25. В следующей главе будут рассмотрены вопросы, касающиеся случайных величин и случайных процессов, которые часто необходимы в задачах моделирования.

§ 31.2. Простой пример дискретного моделирования

Начнем с одного примера с простой схемой моделирования. Пусть речь идет о построении вычислительной машины или какого-нибудь другого крупного электронного прибора. Предположим, что основные части выполнены в виде печатных схем, которые вставляются в большую панель, играющую роль подставки. Предположим далее, что рассматривается только компоновка прибора, а список соединений выходов печатных схем между собой задан, т. е. рассматривается задача инженера-компоновщика.

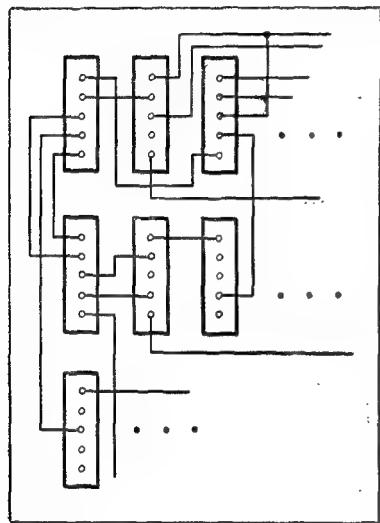


Рис. 31.2-1. Расположение проводов на панели.

В зависимости от того, как расположить отдельные схемы на панели, получится разное расположение проводов. Главное, чем они будут отличаться друг от друга, это общая длина проводов (хотя число проводов будет одним и тем же). По следующим причинам желательно достаточно близко подойти к минимуму длины проводов:

1. Вес проводов может быть серьезным фактором при транспортировке прибора.

2. Стоимость проводов, небольшая в каждом отдельном случае, может стать значительной при массовом производстве.

3. Нарушения схемы (например, случайный контакт между проводами или паразитные емкости) при использовании слишком длинных проводов могут увеличиться.

Для тех, кто незнаком с современной практикой электрических схем, полезно еще одно замечание: условлено прокладывать провода между выходами готовых схем по направлениям север—юг и запад—восток, а не по прямой.

Для тех, кто незнаком с современной практикой электрических схем, полезно еще одно замечание: условлено прокладывать провода между выходами готовых схем по направлениям север—юг и запад—восток, а не по прямой.

Предположим, что мы начинаем с приемлемого, по нашему мнению, расположения схем на панели. На рис. 31.2-1 показана небольшая часть расположения. Нетрудно, хотя, быть может, и утомительно, написать программу для цифровой вычислительной машины, которая возьмет расположение схем на панели и, используя информацию о внутренних соединениях, вычислит расположение и соответствующую длину провода для каждого внутреннего соединения, а также общую длину проводов. Обозначим эту длину L_0 .

Далее можно попытаться поменять две схемы местами. Уменьшит ли это длину проводов? Можно пересчитать все заново, но быстрее иметь другую программу, которая удаляет старые провода и вставляет новые. Таким образом, получается изменение $\Delta L_0 = L_1 - L_0$, где L_1 — новая длина. Если общая длина уменьшилась, то естественно сохранить это изменение и попробовать следующее; если нет, то такого изменения делать не стоит. Таким образом, инженер, komponующий прибор, может использовать машину для моделирования расположения схем на панели. Каждый раз, пробуя изменение общей схемы, он может получить ответ на вопрос, было изменение хорошим или плохим.

Это — один из самых простых примеров моделирования. Здесь нет ни зависимости от времени, ни заслуживающих внимания ошибок вследствие отбрасывания (членов или цифр), ни использования случайных процессов и все очень наглядно: вычисляется, какая получится длина проводов, если принять предложенное расположение схем на панели. Более того, ответ можно выдать в форме, удобной для управления автоматическим укладчиком проводов, и тем самым включить его в общий план автоматизации.

Таким образом, собственно моделирование выполнено. Рассмотрим теперь процесс оптимизации, который на каждом шагу требовал вмешательства человека и при котором терялось очень много машинного времени, если машина быстродействующая. Естественно спросить, почему машина сама не может отбирать и делать пробы. Возможно и верно (а быть может и нет), что инженер может получить ответ при меньшем числе проб; но если дать машине сделать гораздо больше проб, чем хватило бы времени и терпения сделать это инженеру, то время и стоимость работ, а также ее результат могут оказаться удовлетворительными.

Возникает задача — сообщить машине, как делать изменения в расположении схем — изменения, которые, мы надеемся, будут уменьшать общую длину проводов. Можно, конечно, менять местами соседние схемы, сначала соседние по направлению север—юг, затем по направлению запад—восток, и после каждой такой попытки сохранять или отбрасывать перестановку в зависимости от того, уменьшилась или нет общая длина проводов. Проходя, таким образом, многократно по всей панели, пока нельзя будет сделать больше ни одной перестановки, мы постепенно уменьшим общую длину; однако нет никакой уверенности, что в конце мы найдем минимум или приблизимся к нему. В любом случае грубый процесс «перестановки соседей», вероятно, тратит напрасно слишком много машинного времени.

Спросим поэтому себя, как бы мы приступили к этой задаче, если бы делали ее вручную. Способ, который приходит в голову после немногих опытов и наблюдения за собой, следующий. Возьмем произвольную схему и посчитаем, сколько проводов идет от нее на восток, а сколько — на запад. Если больше проводов идет на восток,

то, вероятно, следует рассмотреть передвижение этой схемы в том же направлении. Насколько? Нетрудно понять, что схему нужно передвинуть в такое место, где число проводов, идущих на запад и на восток, одинаково. Повторим затем это же для направления север—юг. Итак, найдется место или по крайней мере область на панели, куда следует поместить выбранную схему. Поищем в этой области вторую схему, обладающую тем свойством, что ее надо было бы передвинуть в район расположения первой. Если такая найдется, то сделаем перестановку. Будем продолжать до тех пор, пока не останется хороших перестановок.

Эту модель можно приблизительно описать так.

Имеется общая тенденция двигать каждую отдельную схему к центру так, чтобы отходящие от нее в каждом направлении провода уравнивались. Конкуренция между схемами заставляет некоторые из них находиться на краях панели. Минимальное расположение схем есть компромисс между этими желаниями.

Вернемся к описанному выше плану действий, который, по-видимому, потребует больше времени и сил на программирование, но меньше машинного времени на счет, чем первый грубый метод. У нас все еще нет уверенности, что на этом пути будет достигнут минимум или получено то же решение, которое находится методом «перестановки соседей». В самом деле, опыт работы такими способами показывает, что если начинать с разных расположений, то получаются существенно разные значения для общей длины проводов. Отсюда вытекает предположение — повторять процесс, начиная с нескольких случайных расположений, чтобы увидеть, какое из них приводит к наименьшей длине. Эта идея оставляет желать лучшего, и следует поискать другие возможности.

Вслепую переставлять одновременно три схемы обычно и не предлагают из-за огромного числа требуемых проб, если количество схем на панели сколько-нибудь велико. Слепые пробы можно заменить более разумным процессом, как, например, тем, который был использован для ускорения перестановки пар, но снова неясно, удастся ли найти нужную тройку. Тогда, может быть, следует попытаться нашим методом переместить первую схему, затем взять ту вторую (или близкую к ней с неуравновешенными проводами) и найти, куда переместить ее, и т. д. в надежде, что какая-нибудь схема в этой цепочке попадет близко к некоторой прежней (не обязательно первой), и таким образом получить цикл замен, которые надо сделать.

В конце процесса перестановки пар будет найден локальный минимум, и мы хотим выйти из него ценой не слишком большого увеличения длины проводов. Отойдя достаточно далеко от района этого минимума, мы можем возобновить простую перестановку пар и спуститься в следующий локальный минимум. Таким образом, мы хотим «встряхнуть» картинку, не слишком увеличивая длину проводов. Если

устроить такой чередующийся процесс, то его можно повторять какое-то время, но после немногих повторений, вероятно, мы окажемся в таком локальном минимуме, который ниже всех близких локальных минимумов. Таким образом, если иметь желание двигаться дальше, то понадобится способ более сильного «встряхивания». Но, прежде чем углубиться в эту интересную задачу, следует спросить себя, так ли в действительности требуется достичь минимума и на каком расстоянии от абсолютного минимума мы согласились бы остановиться, чтобы сберечь машинное (и реальное) время.

К сожалению, мы по-настоящему не имеем представления о том, как далеко находимся от минимума, а можем только, после некоторой практики, догадываться, насколько удастся уменьшить общую длину проводов при заданном объеме вычислений. Это и есть тот реальный вопрос, который встает в большинстве случаев. Известно, что при наличии достаточного машинного времени можно найти минимум путем полного перебора всех комбинаций; но мы не согласны платить такой ценой за минимум, так как выигрыш того не стоит. Решить, как далеко мы хотим идти, можно только в каждом конкретном случае. Но на нынешнем этапе (истории) использования вычислительных машин для оптимизации, вероятно, лучше «перезаниматься» теорией нахождения алгоритмов уменьшения общей длины, чем игнорировать ее; с задачами такого же типа придется встретиться еще много раз, и наверное что-нибудь, чему можно научиться в этом случае, поможет нам в будущем.

Цель этой книги не глубокое проникновение в эту область, а скорее обзор нескольких возможных методов, которые приходят в голову при решении разных задач моделирования. Задача нахождения минимума — это реальная, важная и часто очень трудная задача и, несомненно, одна из тех, внимание к которым будет возрастать в будущем. Уже сейчас существуют такие области, как линейное и динамическое программирование, которые выросли из поиска наилучших или по крайней мере хороших решений.

Здесь, однако, надо сделать одно замечание. В тех ситуациях, которые слишком громоздки для оптимизации в точном смысле, возникает большой соблазн попробовать оптимизировать небольшие части в надежде, что это улучшит положение. Так, в описанном выше примере можно пытаться оптимизировать небольшие блоки схем. Опыт, кажется, показывает, что это в общем плохая тактика; оптимизация малых блоков приводит к большим потерям при окончательном их соединении. Из этого правила, несомненно, есть много исключений, но в целом, кажется, оно соответствует известному принципу системотехники: «Не слишком стараться оптимизировать маленькие тесно связанные части системы, так как при соединении этих частей потеряешь больше, чем выиграл».

§ 31.3. Пример моделирования складских операций

Пример, который будет описан в этом параграфе, относится скорее к тому, что обычно называют практическими приложениями, чем к области научных исследований, но он затрагивает многие проблемы, встающие в обычной научной работе. Из этого реального примера можно извлечь ряд определенных уроков, и отчасти ради них был включен именно этот пример.

На ранней стадии практики ведения складского хозяйства одной из компаний было проведено моделирование на вычислительной машине среднего класса. Изучались главным образом следующие вопросы:

1. Когда заказывать новое оборудование?
2. Сколько оборудования заказывать?
3. Цена отсрочки в выполнении заказа.

До некоторой степени все эти три вопроса находятся под контролем компании. Влияние на третий фактор можно осуществлять многими способами: более быстрым продвижением заказов, передачей заказов по телефону, а не письменно, скорейшей обработкой полученных товаров, главным образом на стадии проверки качества и доставки в действующие хранилища.

В процессе ведения складского хозяйства действуют два противоположных стремления, и именно их столкновение делает задачу интересной. С одной стороны, имея малые запасы, можно уменьшить расходы (и налоги). С другой стороны, желательно избежать «простоев», когда нужного предмета нет на складе, следовательно, нужно стремиться иметь большие запасы. Очевидно, некоторая общая мера стоимости, как, например, оценка предотвращения простоев, даст возможность сравнивать оба обстоятельства и судить о равновесии двух этих противоположных сил.

В принципе моделирование было очень простым. Мы записывали в памяти машины количество наличного товара в начале исследуемого периода. Затем производились действительно наблюдавшиеся выдачи и пополнения склада, которые происходили на протяжении восемнадцати месяцев. После каждой выдачи по оставшемуся количеству проверялось, не пора ли делать новый заказ. Если да, то решалось, сколько заказать, размещались заказы и вычислялось, когда они будут получены. Последняя операция включает случайную величину с распределением Пуассона (здесь на этом можно не останавливаться, поскольку случайные величины будут рассмотрены в следующей главе). Когда наступало время выполнения заказа, его прибавляли к имеющемуся количеству, прибавляя соответствующее число в ячейку, содержащую число наличных предметов. Перед каждой выдачей делалась проверка, осталось ли что-нибудь на складе, и если нет, то регистрировался простой, и надо было следить, через какое время склад пополнялся.

Наконец, через равные промежутки выпечатывалось имеющееся количество вместе со списком простоев и их длительностью.

Прогоняя эту модель для разных правил составления заказов как по времени, так и по объему их, а также изменяя среднее время ожидания заказанных товаров, можно сравнить результаты при различной тактике, и в принципе можно было сказать, что произошло бы по истечении 18 месяцев, если следовать каждому из проверенных правил. Однако в этом рассуждении есть одно слабое место. За реальный период в 18 месяцев были отмечены случайные простои, и потом не отмечалось новых заявок до тех пор, пока товар не поступал. Тем самым моделирование было не вполне точным.

Другой недостаток выясняется, если пытаться узнать не то, что случилось бы, а что произойдет в будущем. Из-за того, что модель выдач и пополнений не построена, а использовалось наблюдавшееся в действительности, не удастся поставить себя в типичную будущую ситуацию. Без сомнения, по частному положению дел в заданные 18 месяцев можно было случайно отдать предпочтение одной формуле, в то время как качество другой выше. Дальнейшее теоретическое исследование могло бы дать лучший результат, но этого сделано не было.

Как же можно оценить эту работу? В большой степени она оценивается потребителями результата. Они считали себя (и действительно были) инженерами-практиками и не слишком верили в теорию. Вычисление того, что произошло бы именно за те 18 месяцев, если бы они придерживались другой тактики (даже когда было указано на недостаточность этого), значило для них гораздо больше, чем теоретические исследования. Мы не будем обсуждать, какой подход лучше, так как у каждого из них есть свои недостатки, а остановимся на том очевидном утверждении, что *бесполезно вычислять результат в такой форме, в которой потребитель не подготовлен его использовать*. То, что вы собираетесь вычислять, должно достаточно хорошо подходить вашему заказчику.

Следует упомянуть еще один момент в этой истории. Хотя нас уверяли, что записи ведутся полно и точно, тем не менее в них были найдены противоречия. Таково общее положение: очень трудно найти точную регистрацию прошлого, особенно если она делается вручную, и не следует принимать на веру чужие слова; чтобы убедиться, что данные соответствуют тому, что о них говорят, надо производить соответствующую их проверку. Большое количество точных данных — вещь чрезвычайно редкая.

§ 31.4. Трехмерные крестики—нолики

Займемся теперь моделированием несколько другого типа, и хотя оно сформулировано в форме игры, принципы его весьма практические. Это — разновидность игры в «крестики—нолики», только в

трехмерном кубе с ребром в четыре клетки. Цель — поставить четыре крестика на одной прямой прежде, чем противнику удастся поставить на одной прямой четыре нолика (предполагается, что читатель знаком с игрой в крестики—нолики на квадрате 3×3 клетки).

В двумерном случае для этой игры известна определенная стратегия, но это, по-видимому, не так для трех измерений. Таким образом, встает задача выработать практический способ выбора следующего хода. В принципе можно перепробовать все возможные продолжения игры, но этот путь, мягко говоря, изнурительный. В этом смысле игра похожа на многие встречающиеся на практике ситуации. В принципе есть возможность полного перебора всех комбинаций, но практически эта возможность недостижима и нам приходится разрабатывать другие методы решения задачи. Преимущество изучения игры, а не практических задач состоит в том, что игра широко известна, проста и четко сформулирована, тогда как многие практические задачи изобилуют мелкими деталями, которые трудно излагать и которые ничего не добавляют к пониманию того, как решать задачу.

Начнем с того, что припишем номера клеткам куба (рис. 31.4-1). Нетрудно видеть, что если четыре крестика стоят на одной прямой,

1	2	3	4	17	18	19	20	33	34	35	36	49	50	51	52
5	6	7	8	21	22	23	24	37	38	39	40	53	54	55	56
9	10	11	12	25	26	27	28	41	42	43	44	57	58	59	60
13	14	15	16	29	30	31	32	45	46	47	48	61	62	63	64

Рис. 31.4-1.

то номера их клеток образуют арифметическую прогрессию. Обратное неверно. Когда игроки делают ход, занимая данное поле, для изучения возможностей выигрыша важна скорее прямая, на которой лежит это поле, чем само поле. Нетрудно видеть, что прямых всего 76. Их можно классифицировать так:

1. По десять прямых на каждой горизонтальной плоскости, из которых две диагональные, — всего 40 прямых.

2. Шестнадцать вертикальных прямых.

3. Шестнадцать наклонных прямых, пересекающих все четыре горизонтальные плоскости (по две из каждой угловой клетки и по одной из каждой боковой).

4. Четыре диагонали куба, начинающиеся в верхнем углу и идущие к диаметрально противоположной вершине куба.

В памяти машины нужно держать и карту куба и список положений на прямых. После каждого хода следует изменить затронутые прямые. Оценка, которая приписывается каждой линии, равна сумме номеров клеток, занятых на этой прямой первым игроком, если на ней нет клеток, занятых вторым игроком, такая же сумма с минусом, если на прямой клетки заняты только вторым игроком, и нуль, если на прямой стоят и крестики и нолики. Еще нужно как-то отмечать совсем пустые прямые.

Посмотрим теперь, как делать ходы в этой игре. Очевидно, следующие правила представляют собой первый шаг на пути к стратегии. Если ход ваш, то:

1. Если у нас есть три крестика на прямой, поставьте четвертый и выиграйте.

2. Если у противника есть три нолика на прямой, займите четвертую клетку, чтобы не дать ему выиграть.

3. Если есть клетка, в которой пересекаются хотя бы две прямые, на каждой из которых стоят по два крестика и нет ноликов, сыграйте в «вилку» и выиграете на следующем ходу.

4. Если у противника есть такая же вилка, то лучше ее заблокировать, а то он выигрывает.

Предположим, что ни одно из этих правил не применимо, — что тогда? Изучим игру несколько глубже. Заметим сначала, что через 16 клеток: 1, 4, 13, 16, 22, 23, 26, 27, 38, 39, 42, 43, 49, 52, 61, 64 — проходит по семь прямых, тогда как через все остальные клетки только по четыре. Из этого следует, что при прочих равных условиях, играя на этих клетках, получаешь больше шансов впоследствии найти выигрывающую комбинацию. Ситуация, в которой мы теперь оказались, характерна для многих подобных задач: есть определенные алгоритмы, как действовать в некоторых случаях, и необходимо придумать, как действовать в остальных.

Поскольку главная цель примера — проиллюстрировать этот принцип, интерес к этой игре, возможно, оправдывается следующими несколькими замечаниями, хотя у нас и нет полной стратегии.

Дальнейшее изучение игры показывает, что есть такое преобразование — инверсия куба самого в себя, — которое оставляет инвариантными все прямые и меняет местами восемь вершин с восемью центральными клетками (рис. 31.4-2). Таким образом, любая стратегия для центральных клеток эквивалентна такой же стратегии для вершин.

В этом месте, вероятно, необходимо сыграть достаточное число партий, чтобы представлять себе ход игры. Судя по практике автора игра разворачивается примерно так. Оба игрока начинают играть главным образом на «лучших» клетках, через которые проходит семь прямых, но вскоре один из них начинает последовательность форсирующих ходов, ставя по три креста на прямую, на что другой

может отвечать только блокировкой в очевидных местах. Первый игрок надеется построить вилку, против которой нет защиты, и, когда он выбирает свои форсирующие ходы, он отчасти имеет это в виду. Он должен также следить, чтобы защитный ход противника не передал неожиданно ему инициативу, предоставляя возможность форсирующего хода. Если первый атакующий не сумел добиться

22	21	24	23	6	5	8	7	54	53	56	55	38	37	40	39
18	17	20	19	2	1	4	3	50	49	52	51	34	33	36	35
30	29	32	31	14	13	16	15	62	61	64	63	46	45	48	47
26	25	28	27	10	9	12	11	58	57	60	59	42	41	44	43

Рис. 31.4-2.

победы и вынужден отдать инициативу, то обычно второй игрок атакует и доводит дело до победы. Кажется невозможным сколько-нибудь долго вести чисто защитную игру против агрессивного противника и не проиграть. Если пытаться превратить эти неопределенные рассуждения в стратегию, то легко обнаружить, что на разных стадиях игры требуются совсем разные стратегии. В начале игры — это попытки заполнить лучшие клетки, защититься против ранних атак, а также помешать противнику поставить слишком много ноликов на какой-нибудь плоскости, где нет наших крестиков, блокирующих комбинации, которые он может там развить.

Придется, далее, придумать способ определения подходящего момента для перехода к атаке, помня о том, что если начать слишком рано, то придется отдать инициативу и кончить поражением, а если опоздать, то противник может атаковать и пробиться к победе.

Ясно, что необходимо включить в каждый форсирующий ход анализ ответа противника и ситуацию, которая возникнет, чтобы увидеть, не потеряем ли мы инициативу. Следует также стремиться выбрать такой форсирующий ход, при котором мы ходим на лучшую клетку, а противник — нет. При сколько-нибудь глубоком изучении хорошо было бы не только учитывать расположение клеток и положение на прямых, но еще и разработать стратегию для плоскостей. Опыт партий, иггранных людьми, заставляет предположить, что машина с такой программой может, вероятно, играть вполне хорошо благодаря своей способности каждый раз тщательно проверять алгоритмическую часть и не зевать комбинации, что обычно делают люди. С другой стороны, сделать программу, просчитывающую длинные комбинации, по-видимому, трудно.

Следующие упражнения дают некоторые комбинаторные результаты, найденные при подробном исследовании игры; они характерны как частные результаты, которые можно найти в сложной задаче. Методы их решения в тексте не указаны.

Упражнения

31.4-1. Показать, что 16 надлежащим образом расставленных крестиков могут блокировать 73 прямые (кроме трех главных диагоналей). (Один из ответов: 1, 8, 10, 15, 19, 22, 28, 29, 36, 37, 43, 46, 50, 55, 57, 64.)

31.4-2. Показать, что если модифицировать правила игры так, чтобы первые восемь ходов должны были делаться только в центральный $2 \times 2 \times 2$ куб (или в восемь вершин), то первый игрок выигрывает.

§ 31.5. Общие замечания о дискретном моделировании

Немногие приведенные примеры показывают, что обычно трудной частью дискретного моделирования является не нахождение способа устроить модель, хотя это и может практически оказаться утомительным упражнением в программировании для конкретной машины, а нахождение алгоритма решения тесно примыкающей задачи оптимизации. Обычно в дискретных случаях существует известный метод полного перебора всех вариантов, который так дорог, что его нельзя использовать. Тем самым приходится применять эвристические методы и переходить в область узнавания образов и думающих машин.

Когда что-нибудь в этом направлении применяется к частной задаче, обычно оказывается, что в немногих случаях есть определенный алгоритм, вроде того, который был указан для крестиков—ноликов, но в большинстве случаев ситуация менее определена. Когда ни алгоритмические, ни эвристические правила не дают определенного хода, все же, как в случаях крестиков—ноликов, какой-то ход должен быть выбран.

Здесь стоит обратиться к понятию случайного хода или случайного хода из данного класса. Если ходы выписаны в каком-нибудь порядке, то можно, конечно, брать всегда первый из выписанных ходов. Это не так хорошо, как делать случайный ход, по нескольким причинам, в числе которых следующие:

1. Если всегда выбирается определенный ход, противник может постепенно изучить игру, найти слабое место и воспользоваться им, тогда как при использовании случайных ходов это очень трудно.

2. Неизвестно почему, но систематический выбор в течение большого времени может дать плохой эффект, и мы чувствуем интуитивно, что случайный ход принесет нам в среднем меньше вреда.

Итак, мы пришли к необходимости заняться «случайностью», что и будет сделано в следующей главе.

В этом месте, возможно, следует привести несколько слов предостережения. Часто после изучения оказывается, что есть другие формулировки исследуемой ситуации и что одни из них много более удобны для моделирования на машине, чем другие. Поэтому читателю не следует торопиться влезать в детали моделирования; в особенности следует убедиться, что предполагаемое моделирование ответит на те вопросы, на которые надо ответить.

§ 31.6. Непрерывное моделирование

В большинстве случаев непрерывное моделирование включает решение одного или нескольких дифференциальных уравнений. В таких моделях часто употребляются функции с ограниченным спектром, так как моделируемые электронные схемы могут пропускать обычно ограниченную полосу частот. Тем самым оказывается, что к этому вопросу частично относятся главы с 21 по 25, особенно глава 25.

Однако детали разных случаев моделирования часто очень различны, и прежде чем начинать программировать численное решение, необходимо тщательно изучить предполагаемую модель. Небрежность в вопросе выбора достаточно малого шага, приводящая к нежелательному переучитыванию частот, есть лишний пример того, как спешка на стадии планирования может сказаться на полученных результатах.

К большим задачам следует приступать с осторожностью и не торопясь. Если учитывать цену всего машинного времени, потраченного на задачу, то ясно, что не надо жалеть времени и сил на стадии планирования.

Одной из главных целей этой книги было изложение идей и соображений относительно методов нахождения вида формул, требуемых конкретной ситуацией. Автор надеется, что книга даст читателю умение и уверенность в себе, необходимые для того, чтобы вынести долгие месяцы тщательной отработки и проверки подходящих формул для моделирования.

В непрерывных моделях часто встречается моделирование шума, и приведенные в следующей главе методы получения случайных чисел помогут приступить к построению шума с нужными свойствами спектра, автокорреляции и т. д.

От конкретной ситуации зависит так много, что невозможно рекомендовать какой-нибудь определенный набор формул, и мы вынуждены быть неопределенными. Почти единственным общим правилом является то, что, прежде чем входить в детали, необходимо осознать общий план моделирования и связи его со всей задачей. С этой стороны вопрос освещается главным образом в главе $N+1$.

ГЛАВА 32

СЛУЧАЙНЫЕ ЧИСЛА И МЕТОДЫ МОНТЕ-КАРЛО

§ 32.1. Понятие случайного числа

В математике и в статистике существует точно определенное понятие случайного процесса. Понятие случайного числа не столь просто. Случайные числа суть результат случайного процесса. Но можно ли считать случайной уже записанную последовательность чисел? Ведь раз она записана, это уже вполне предсказуемая последовательность. Очевидно, прежде чем заниматься вопросами использования случайных чисел, над этим надо еще подумать.

Под случайным целым числом между 0 и 9, вообще говоря, можно понимать цифру, выбранную из совокупности, когда все цифры имеют равную вероятность быть выбранными. Совсем другое дело — случайное число между 0 и 1. Почти для всех таких чисел перечисление цифр заняло бы бесконечное время. Устроить случайную величину в машине вовсе не означает случайную величину в математическом смысле. Что же это означает? Ну, например, это могло бы означать, что мы собираемся из совокупности всех восьми-, десяти- или двенадцатизначных чисел выбрать одно некоторым «равновероятным» способом. На самом деле мы пойдем на дальнейший компромисс и будем выбирать только из части этой совокупности чисел. Что имеется в виду, когда собираются устроить последовательность случайных чисел? Обычно мы имеем в виду просматривать некоторую часть всей совокупности представимых в данной машине чисел, брать их по одному и не повторяться, пока не исчерпаем их все. Это можно было бы назвать выборкой без повторений.

Но если мы собираемся получить целую последовательность чисел, то в дополнение к равновероятности понадобятся дальнейшие проверки. Какую последовательность требуется получить? Если ее члены будут монотонно возрастать, то ее не хотелось бы считать случайной, хотя в некотором смысле она столь же случайна, как и любая другая конкретная последовательность. Очевидно, требуются еще какие-то свойства.

Первым делом, говорится, что последовательность должна быть построена случайным способом. Но что это такое? Ведь всякий порядок столь же случаен, как и любой другой. Приходится сказать, что не должна быть видна какая-нибудь закономерность в способе, которым выбрана последовательность. Один из законов, для которых можно устроить проверку, — закон корреляции числа и следующего за ним; можно ли предсказать следующее число по предыдущему? Таким образом, если x_i — случайные числа, то для нашей

последовательности должно выполняться равенство

$$\sum_i \left(x_i - \frac{1}{2}\right) \left(x_{i+1} - \frac{1}{2}\right) = 0,$$

где взято среднее $\bar{x} = 1/2$. Имея достаточно много времени, можно придумать очень много возможных проверок, так много, что не все их удастся применить.

В этом вопросе необходимо стать на практическую точку зрения: если по отношению к данному частному применению закон незаметен, то для данного применения числа случайны, и мы должны этим удовлетвориться. В самом деле, с практической точки зрения, быть может, и не требуется иметь действительно случайную последовательность, а предпочтительнее более гладкое распределение, исчерпывающее часть совокупности без повторений.

Способ, которым обычно используются случайные последовательности во многих приложениях, состоит в том, чтобы пытаться оценить статистику большой совокупности событий по малой выборке. Поэтому хочется получить «сверхтипичную» выборку так, чтобы для малой выборки иметь в реальном случае стабильность, как для большой. Хочется иметь более гладкое распределение, чем обычно, чтобы было не слишком много больших выбросов — ровно столько, сколько надо, чтобы не впасть в противоположную ошибку слишком однородной совокупности. Мы хотим, если это возможно, устроить выровненную, очищенную и проверенную последовательность случайных чисел.

Высказанные замечания не универсальны; они неверны, например, если требуется оценить границы изменения какой-нибудь величины.

§ 32.2. Генерирование случайных чисел в машине, работающей в двоичной системе

Наиболее широко употребляется следующий метод получения случайных чисел:

$$x_{n+1} = r x_n \text{ (по модулю два в степени длины машинного слова)}$$

Таким образом, используется одно умножение на каждое число и младшие разряды произведения берутся в качестве следующего случайного числа. Основной вопрос состоит в том, что взять за r и как выбрать x_0 .

Ответ на эти вопросы и доказательство того, что данный метод до повторения дает длинные последовательности случайных чисел, основываются на некоторых результатах теории чисел. Для облегчения вывода *) (цель которого — продемонстрировать метод и резуль-

*) Заимствованного у М. Леви.

таты) рассмотрим задачу получения случайных чисел на машине, работающей с двоичными числами. Результаты для десятичной машины приведены в следующем параграфе.

Обычная в теории чисел запись

$$x \equiv a \pmod{m}$$

означает, что $x - a$ делится на m .

Если наша машина k -разрядная и берутся последние k разрядов произведения (удвоенной длины) ρx_n , то

$$\left. \begin{aligned} x_0 &= a, \\ x_{n+1} &\equiv \rho x_n \pmod{2^k} \quad (k \geq 3). \end{aligned} \right\} \quad (32.2-1)$$

Если a делится на 2, то последовательность будет эквивалентна такой:

$$y_0 = \frac{a}{2}, \quad y_{n+1} \equiv \rho y_n \pmod{2^{k-1}},$$

которая на самом деле соответствует более короткому машинному слову, и значит, здесь не используются все возможности нашей машины. Поэтому возьмем в качестве a нечетное число.

Аналогично ρ должно быть нечетным, так как если бы оно было четным, то

$$x_{k+1} = \rho^{k+1} a = 0$$

и с этого места мы бы имели тривиальную последовательность нулей.

Далее, все нечетные числа ρ можно записать одним из способов:

$$8t - 3, \quad 8t - 1, \quad 8t + 1, \quad 8t + 3.$$

Теорема 1. Если $\rho = 8t \pm 1$, то

$$\rho^{2^{k-3}} \equiv 1 \pmod{2^k}, \quad (32.2-2)$$

т. е. порядок ρ не больше, чем 2^{k-3} , и является делителем числа 2^{k-3} .

Доказательство. Утверждение теоремы эквивалентно тому, что разность $\rho^{2^{k-3}} - 1$ делится на 2^k . Так как

$$a^2 - 1 = (a + 1)(a - 1),$$

то

$$\rho^{2^{k-3}} - 1 = (\rho^{2^{k-4}} + 1)(\rho^{2^{k-5}} + 1) \dots (\rho + 1)(\rho - 1). \quad (32.2-3)$$

Для каждой скобки ($i \geq 1$)

$$\rho^{2^i} + 1 = (8t \pm 1)^{2^i} + 1 = 1 + (1 \pm 8t)^{2^i} = (1 + 1) + \sum_{k=1}^{2^i} (-1)^k C_{2^i}^k t^k 8^k,$$

отсюда 2 делит $\rho^{2^i} + 1$ и из (32.2-3) следует, что 2^{k-4} делит

$$(\rho^{2^{k-4}} + 1)(\rho^{2^{k-5}} + 1) \dots (\rho^2 + 1). \quad (32.2-4)$$

Далее,

$$(\rho + 1)(\rho - 1) = (\rho^2 - 1) = (8t \pm 1)^2 - 1 = 16(4t^2 \pm t),$$

поэтому $(\rho + 1)(\rho - 1)$ делится на 2^4 и из (32.2-3), учитывая (32.2-4), получаем, что $2^4 \cdot 2^{k-4} = 2^k$ делит $(\rho^{2^{k-3}} - 1)$.

Теорема 2. Если $\rho = 8t \pm 3$, то

$$\rho^{2^{k-3}} \not\equiv 1 \pmod{2^k}, \quad \rho^{2^{k-2}} \equiv 1 \pmod{2^k},$$

т. е. порядок ρ равен 2^{k-2} .

Доказательство. Используя (32.2-3) в этом случае, получаем

$$\rho^{2^i} + 1 = 1 + (3 + 8t)^{2^i} = 1 + 3^{2^i} + \sum_{k=1}^{2^i} (\pm 1)^k C_{2^i}^k t^k 8^k 3^{(2^i-k)}.$$

Но

$$(1 + 3^{2^i}) = 1 + (4 - 1)^{2^i} = 1 + (1 - 4)^{2^i} = 1 + 1 + \sum_{k=1}^{2^i} C_{2^i}^k (-4)^k.$$

Отсюда

$$\rho^{2^i} + 1 = 2 + \sum_{k=1}^{2^i} C_{2^i}^k [3^{2^i-k} t^k \cdot 2^k + (-1)^k] 4^k$$

и $\rho^{2^i} + 1$ делится на 2, но не делится на 4.

Следовательно, из равенства (32.2-3) получаем, что 2^{k-4} делит $(\rho^{2^{k-4}} + 1)(\rho^{2^{k-5}} + 1) \dots (\rho^2 + 1)$, но 2^{k-3} не делит это произведение.

Заметим, что

$$(\rho + 1)(\rho - 1) = 8(8t^2 \pm 6t + 1) = 8 \times (\text{нечетное число}).$$

Таким образом, в любом случае $(\rho + 1)(\rho - 1)$ делится на 2^3 и не делится на 2^4 . Следовательно, $(\rho^{2^{k-3}} - 1)$ делится на 2^{k-1} , но не делится на 2^k . Иными словами,

$$\rho^{2^{k-3}} \not\equiv 1 \pmod{2^k}.$$

Запишем теперь

$$\rho^{2^{k-1}} - 1 = (\rho^{2^{k-2}} + 1)(\rho^{2^{k-2}} - 1),$$

причем 2 делит $\rho^{2^{k-3}} + 1$ и 2^{k-1} делит $\rho^{2^{k-3}} - 1$, так что 2^k делит их произведение и

$$\rho^{2^{k-2}} \equiv 1 \pmod{2^k}.$$

Теорема 3. Если $p = 8t - 3$, то последовательность $x_0, x_1, \dots, x_{(2^{k-2}-1)}$, порожденная формулой (32.2-1), есть некоторая перестановка последовательности

$$1, 5, 9, \dots, 2^k - 3, \text{ если } a \equiv 1 \pmod{4},$$

или

$$3, 7, 11, \dots, 2^k - 1, \text{ если } a \equiv 3 \pmod{4}.$$

Доказательство. Рассмотрим значение (ap^n) , $n = 0, 1, \dots, 2^{2^{k-2}} - 1$. Разность между последовательными членами

$$ap^{n+1} - ap^n = ap^n(p - 1) = ap^n(8t - 4)$$

делится на 4. Но уже известно (теорема 2), что порядок числа p равен 2^{k-2} . Следовательно, получаются 2^{k-2} различных членов, разности которых делятся на 4, из чего следует утверждение теоремы.

Итак, в зависимости от выбора a и при условии, что $p = 8t - 3$ (для любого t), можно получить перестановку одной из двух последовательностей, указанных в теореме 3. При выводе этого факта x_n считались целыми числами; при использовании их слева ставится двоичная запятая и используются числа

$$x_n \cdot 2^{-k}.$$

Но значение t нехорошо выбирать совсем произвольно. Например значение $t = 1$ дает $p = 5$, и если где-нибудь встретится маленькое x_n , скажем 10^{-k} , то за ним получится длинная последовательность постепенно возрастающих чисел. Во избежание этих неприятностей можно было бы выбрать такое t , чтобы старшие разряды $p \cdot 2^{-k}$ были или 001, или 010. Подробного изучения преимуществ выбора различных значений t для получения различных перестановок тех же самых чисел, по-видимому, нет.

Для данного $p = (8t - 3)$ мы получили два цикла, каждый из которых исчерпывает четверть всех имеющихся в машине чисел. Оставшаяся половина чисел входит в меньшие циклы, зависящие от выбора a .

Если в последовательностях теоремы 3 опустить два последних двоичных разряда, то полученная последовательность будет перестановкой полного цикла длины 2^{k-2} : $0, 1, 2, \dots, 2^{k-1} - 1$. Если не опустить два последних разряда, то последний разряд всегда будет 1 и, конечно, не будет случайным. Переходя от правых разрядов к левым и проверяя разные встречающиеся разряды, находим все меньше и меньше закономерностей. Поэтому принято не рассчитывать на случайность нескольких последних разрядов.

§ 32.3. Генерирование случайных чисел на десятичной машине

Основные результаты в этой области изложены в статье Мошмана*), где приводятся формула для s -значной десятичной машины ($s \geq 4$)

$$p = 7^{4s+1},$$

которая имеет период длины $5 \cdot 10^{s-3}$, а также результаты проверки ее на одиннадцатизначной машине. Другой выбор**) $p = 76\,768\,779\,754\,638\,671\,877$ (приведенное по модулю 10^5) дает максимальные периоды.

Есть и дальнейшие исследования, использующие более сложные формулы. Формулу для двоичной машины.

$$x_{n+1} = (2^a + 1)x_n + c \pmod{2^{35}}$$

можно найти в статье Ротенберга***).

Изучались также формулы вида

$$x_{n+1} = \alpha x_n + \beta x_{n-1}.$$

Ясно, что проблема получения случайной величины на машине быстро развивается и не может здесь быть исчерпана. Мы оставляем эту тему для дальнейших исследований.

§ 32.4. Другие распределения

Зная, как получать равномерно распределенные случайные числа, можно генерировать другие распределения. Известно много различных способов. Рассмотрим некоторые из них.

Предположим, что требуется распределение $f(y)$; как получить его из равномерного? Один из способов — приравнять две функции распределения, равномерную для переменной x и желаемую для y :

$$\int_0^x 1 \cdot dx = \int_0^y f(y) dy = F(y) = x.$$

Таким образом, если можно найти функцию, обратную к $F(y)$, то имеем

$$y = F^{-1}(F(y)) = F^{-1}(x).$$

В принципе, это все, что требуется.

*) Jack Moshaman, The Generation of Pseudo-random Numbers on a Decimal Machine, J. Assoc. Computing Machinery, vol. 1, pp. 88—91, 1954.

**) Принадлежащий Е. Н. Гилберту.

***) A. Rotenberg A. New Pseudo-random Number Generator, J. Assoc. Computer Machinery, vol. 7, pp. 75—77, 1960. См. также J. Certain, On Sequences of Pseudo-random Numbers of Maximal Length, J. Assoc. Computer Machinery, vol. 5, pp. 353—356, 1958.

Простым примером этого способа является экспоненциальное распределение

$$f(y) = e^{-y}, \quad x = \int_0^y e^{-y} dy = 1 - e^{-y}, \quad (32.4-1)$$

или

$$e^{-y} = 1 - x.$$

На практике можно заменить случайную величину $1 - x_i$ величиной x_i , так что $e^{-y_i} = x_i$ или

$$y_i = -\ln x_i.$$

Эта случайная величина употреблялась с большим успехом.

Если применить этот метод к нормальному распределению $\frac{1}{\sqrt{2\pi}} e^{-y^2}$, то придется приближать функцию, обратную erf (например, по Гастингсу [13], стр. 191). Но опыт показывает, что для обратных функций распределения приближения иногда ведут к серьезным ошибкам.

Вместо того чтобы брать обратную функцию, можно воспользоваться тем хорошо известным фактом, что сумма сравнительно небольшого числа случайных величин с любым распределением обычно дает очень хорошее приближение к нормальному распределению. Для равномерного распределения требуется сумма около десяти чисел, чтобы получить почти нормальное распределение. В десятичной машине принято брать двенадцать чисел, чтобы избежать возможного взаимодействия между механизмом генерирования и процессом сложения. Сумма не дает в среднем нуля, и поэтому нужно вычитать необходимую величину: 5, если складывались десять чисел, и 6, если складываемых было двенадцать. Используя результаты § 2. 9, получаем дисперсию

$$\sigma^2 = \frac{n}{12} \quad (n = 10, 12).$$

Выбирая n равным 12, получаем дисперсию 1, что является вторым доводом в пользу $n = 12$.

В случае пуассоновского распределения успешно употребляется следующий способ. Находим такое k , чтобы для x_i из равномерного распределения выполнялись неравенства

$$x_1 x_2 \dots x_k \geq e^{-m}, \quad x_1 x_2 \dots x_{k+1} < e^{-m},$$

где m — среднее пуассоновского распределения. Это k и есть нужное число.

Если m мало, эта процедура экономична, но при достаточно больших m она занимает много времени и не слишком хороша. Впрочем, если m достаточно велико, то распределение Пуассона довольно точно приближается нормальным.

§ 32.5. Метод Монте-Карло

Мы привели некоторые из многих известных способов получения случайных величин с разными распределениями и теперь приступаем к описанию возможных способов их использования. Название «методы Монте-Карло» для методов, систематически использующих случайную величину, восходит к последним годам второй мировой войны, когда фон Нейман и Улам использовали случайные числа для моделирования поведения нейтронов. Комбинированное использование машин с людьми для моделирования случайного процесса вызовов на телефонной станции было осуществлено не позже 1926 года, хотя в те дни это называлось (а кое-где и теперь называется) «бросанием» по аналогии с бросанием кости для получения случайных чисел. Но сама идея восходит по меньшей мере к французскому естествоиспытателю Бюффону, который заметил, что если иглолку бросать случайным образом на разлинованную плоскость, то вероятность того, что иглолка пересечется с какой-нибудь линией, связана с числом π .

Установим этот факт. Пусть дано семейство равноотстоящих параллельных прямых линий с расстоянием 1 между ними. Пусть длина

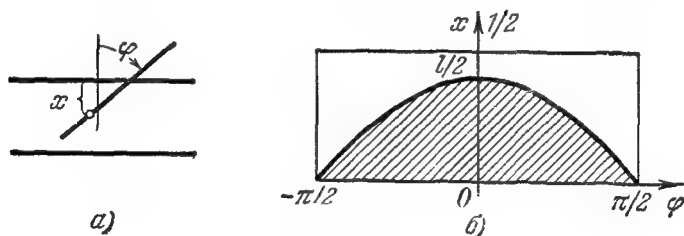


Рис. 32.5-1.

иглолки $l < 1$. Когда иглолку бросают, центр ее может упасть от прямой на любом расстоянии от 0 до $1/2$. Пусть это будет переменная x . Рассмотрим, далее, угол φ наклона иглолки к прямым (рис. 32.5-1, а). Переменные x и φ можно считать случайными и независимыми.

Условие пересечения иглолки с одной из линий есть

$$x < \frac{l}{2} \cos \varphi \left(-\frac{\pi}{2} < \varphi < \frac{\pi}{2} \right).$$

На рис. 32.5-1, б заштрихована область, ограниченная кривой $x = \frac{l}{2} \cos \varphi$, в которую должна попадать точка с координатами (φ, x) .

Отношение площади этой области к площади всего прямоугольника возможных точек есть вероятность пересечения. Это отношение равно

$$P = \frac{\int_{-\pi/2}^{\pi/2} \frac{l}{2} \cos \varphi d\varphi}{\frac{1}{2} \pi} = \frac{l}{\pi} \int_{-\pi/2}^{\pi/2} \cos \varphi d\varphi = \frac{2l}{\pi}.$$

Таким образом, можно рассматривать рассуждения Бюффона как определение значения π методом Монте-Карло. Если бы про π было известно только, что оно лежит между 1 и 10, это был бы прекрасный способ обнаружить, что π примерно равно 3. Затратив некоторый труд и сделав много бросаний, мы могли бы получить π примерно 3,1. Но, чтобы получить большую точность, нужно было бы проводить прямые, измерять длину иголки, разбирать случаи спорного пересечения и т. д. с точностью большей, чем в наших силах, либо производить невероятно большое число бросаний. И это есть общий факт. Метод Монте-Карло, быть может, лучше других на предварительной стадии, когда он помогает получить общее представление о ситуации, но если требуется получить точные результаты, ценность его значительно меньше.

Иллюстрируя это замечание, приведем пример из практики автора. Когда только начали появляться вычислительные машины, была предложена задача, которую в первоначальной аналитической форме было бы трудно решить и на самых быстрых современных машинах, но оказалось, что на самом деле это задача о движении иона в электрическом поле в газе. Моделирование по методу Монте-Карло с 10 000 частиц дало график распределения скорости вдоль поля и перпендикулярно к нему. После того как физик перестал жаловаться на низкую точность, он сказал нечто вроде: «Хм... Это похоже на эллиптическое распределение Максвелла, только слегка сдвинутое. Хм...». И это дало ему ключ к аналитическому решению задачи, — только так и были использованы численные результаты моделирования. К сожалению, на самом деле метод Монте-Карло в большинстве случаев используется только тогда, когда прочие методы уже исчерпаны, а в этих обстоятельствах он теряет некоторые свои преимущества.

Упражнение 32.5-1. Описать программу, моделирующую задачу Бюффона на цифровой машине.

§ 32.6. Еще одна иллюстрация метода Монте-Карло

Обычно реальная задача включает элемент случайности, что наталкивает на использование случайных чисел в моделирующем процессе. Но иногда, как в примере с иглой Бюффона, необходимо найти

другую формулировку задачи, чтобы она уже содержала случайный элемент. (Некоторые дают имя «Монте-Карло» только таким случаям.)

Приведем еще следующий пример использования статистического подхода.

В психологическом эксперименте группе из пяти человек, сидящих за столом, была дана задача для совместного решения. Изучалось в основном поведение их в зависимости от изменения каналов связи между ними. Оказалось, что некоторая организация этой связи дала лучший результат, чем другие.

Однако при этом не было ответа на вопрос: «насколько хороши были системы связи?» То, что измерялось, было их взаимное соотношение. Естественно задать вопрос: «А если бы люди действовали случайным образом, каков был бы результат?» В некотором смысле исследование случайного поведения дает абсолютную меру успеха организации их работы. Такой эксперимент также яснее отражает структурные эффекты, получающиеся благодаря наличию тех или иных каналов связи.

Таким образом, иногда статистический подход, не являясь никоим образом частью модели исследуемой системы, может тем не менее внести вклад в понимание работы модели. Не обязательно предполагать, что люди действуют случайным образом, но ответ на вопрос: «Что было бы, если бы они действовали случайно?» — проясняет наблюдаемую картину.

§ 32.7. Метод жулика

Ввиду популярности красочных слов «Монте-Карло» мы приняли другое живописное выражение «метод жулика» *) для описания использования в вычислениях по Монте-Карло коррелированных величин. Удачное использование коррелированных (особенно отрицательно коррелированных) величин может сильно уменьшить объем вычислений при моделировании по Монте-Карло.

Как уже часто случалось, обсуждая различные вычислительные вопросы, мы быстро углубились в статистику и вынуждены признать, что эта тема выходит за рамки вводного курса по методам вычислений. Но выгода таких методов в случаях, когда их можно заставить работать, слишком велика, чтобы не сказать о них хоть несколько слов. Приведем лишь один пример. Хоммерсли и Мортон **) описывают применение отрицательно коррелированных величин в экспери-

*) Принадлежит, по-видимому, проф. Тьюки и используется для обозначения любого метода ускорения вычислений переходом к эквивалентной задаче.

**) J. M. Hammersley and K. W. Morton, A New Monte Carlo Technique: Antithetic Variates, Proc. Cambridge Phil. Soc., vol. 52, pt. 3, pp. 419—457, 1956.

менте Бюффона. Они сначала берут две жестко скрепленные крестом иголки и показывают, что получается выигрыш в числе необходимых бросаний в 12,2 раза (одно бросание и регистрация пересечения креста засчитываются за два бросания с регистрацией простой иглы), т. е. можно ожидать, что для получения той же точности с крестом придется сделать в 12,2 раза меньше бросаний. Когда они перешли к трем иголкам, пересекающимся под равными углами (и считали одно бросание за три), был достигнут выигрыш в 44,3 раза; для четырех выигрыш был в 107,2 раза. Тот же результат можно получить, скрепив иголки в виде правильного треугольника. Далее, они показывают, что с помощью некоторых методов статистики можно получить еще больший выигрыш. Итак, прежде чем начинать вычисления по Монте-Карло, советуем получить консультацию у хорошего статистика, чтобы посмотреть, чем он может помочь.

ГЛАВА N + 1

ИСКУССТВО ВЫЧИСЛЯТЬ ДЛЯ ИНЖЕНЕРОВ И УЧЕНЫХ

§ N + 1.1. Важность вопроса

В книгу по методам вычислений не принято включать главу на неопределенную, общую тему: как подходить к задаче и решать ее, если решение требует использования вычислительных машин. Сам заголовок этой главы двусмыслен — его можно понимать как адресованный человеку, вычисляющему для инженеров и ученых, так и к ним самим; на самом деле имеется в виду и то и другое.

Не следует относиться несерьезно к этой теме только из-за того, что в некоторых местах излагается скорее личное мнение, чем установленные факты. Автору она представляется более важной, чем многие конкретные результаты, изложенные в других частях книги. В настоящее время вся тема представляет собой скорее искусство, нежели науку, но это положение быстро меняется, так как само наличие вычислительных машин сделало возможной механизацию некоторых процессов, про которые когда-то полагали, что они требуют человеческой мысли, и теперь в этой важной области продолжаются активные поиски.

Чем больше мы узнаем о том, как мы решаем задачи, тем большую часть работы можно будет переложить на машины. Искусство решения задач на вычислительных машинах интересно и само по себе; оно может помочь во многих ситуациях и сильно увеличить ценность машинных вычислений.

Большинство ученых избегают заниматься исследованием того, что, вообще говоря, можно назвать процессом творческого мышления, но.

есть несколько замечательных исключений. «Метод» Архимеда, пожалуй, один из самых ранних примеров таких исключений среди математиков, а теперь имеется еще современная классическая книга Поля «Как решать задачу» [36]. Однако оба этих автора занимались решением точно сформулированных задач, тогда как нам интересно, что делать, когда задача поставлена нечетко и столь же неопределенны условия, которым должны удовлетворять результаты. Девиз этой книги

Цель расчетов — не числа, а понимание

показывает, насколько широка наша область исследований.

§ N + 1.2. Что мы собираемся делать с ответом?

Поля в своей книге «Как решать задачу» [36] показывает, как важно понять условия. Автор этой книги в результате многолетней практики вычислений на заказчика пришел к убеждению, что обычно первым делом следует подумать: «А что мы собираемся делать с ответом?» Будут ли вычисленные величины действительно отвечать на вопрос, который нам задан? Все ли они нам нужны? Может быть, нужны еще какие-нибудь? Может быть, что-нибудь другое дает лучшие основания для понимания?

Чтобы ответить на некоторые из этих вопросов, можно представить себе типичный листок с ответом и проверить его на полезность. Гораздо чаще, чем можно было бы думать, результаты, которых от нас требуют, не соответствуют нуждам задуманного исследования.

Например, первоначальным требованием могло быть решение системы уравнений. Иногда это и все, что вычисления могут дать, но во многих случаях пониманию исследуемой ситуации могут способствовать другие вещи, например трудность решения. Далее, что должно быть мерой точности: точность по неизвестным, по правым частям или еще что-нибудь? Необходимо ли решать эту систему? И, наконец, нельзя ли все лучше понять другим способом?

Прежде чем продолжать в таком же духе, заметим, что и нельзя ожидать от заказчика, чтобы он точно знал, что он хочет. На многих стадиях исследовательской работы не знать в точности, что ты ищешь, вполне естественно. В самом деле, можно сказать: «Если исследователь знает, что он делает, то этого не надо было делать». В некотором смысле, если получается в точности ожидаемый результат, то мы не узнаем ничего нового, хотя может возрасти уверенность в чем-то *).

Как бы тривиально и очевидно это не звучало, повторим еще раз: важно понимать, что вы хотите узнать. Гораздо реже понимают, что

*) П. Дебай: «Если задача ясно поставлена, то для физика она не представляет больше интереса».

работу надо специально планировать так, чтобы увеличить шансы заметить что-нибудь необычное. Если можно включить в процесс счета дополнительные побочные проверки исследуемой модели, то ради этого стоит потратить еще немного машинного времени. Многие великие открытия были сделаны в результате случайного наблюдения, важность которого понял подготовленный исследователь. Так, полезно, кроме самого минимума, выпечатывать еще несколько разумно выбранных чисел, несмотря на то, что это дополнительно загружает выходные устройства.

Итак, общее доброе правило (хотя и не без исключений): приступая к вычислительной задаче, попытаться ответить на вопрос: *«Что мы собираемся делать с ответом?»* Активность и воображение могут дать многое для всего исследования уже на этой стадии, в то время как лень и скука могут помешать возникновению какого бы то ни было понимания, расходуя многие часы счета для получения очевидных числовых результатов.

Одна из наиболее частых ошибок — потребовать вывода слишком многих величин, особенно в задачах со многими параметрами. В таких случаях обычно нужнее всего хороший статистик, знакомый с *теорией планирования экспериментов*. Очень часто он может так изменить постановку исследований, что понадобится обработка лишь небольшой части первоначального числа случаев. Вывод полного объема вычисленных величин может задушить всякое понимание.

§ N + 1.3. Что мы знаем?

Теперь, поняв, что мы намерены извлечь из вычислений, зададимся вопросом: *«Что нам известно?»* Какой информацией мы располагаем? Каковы входные данные? Включили ли мы в них все, что нам известно? Например, если известно, что решение проходит через начало координат, учтен ли этот факт во входных данных?

Мы снова напоминаем читателю, что Поля подчеркивает важность понимания задачи, но он имеет в виду математически поставленные задачи и предполагает, что условия задачи заданы полностью. Для приложений математики это, конечно, не так; неверно это и в собственно математических исследованиях. Часто дополнительные проверки изучаемой ситуации приносят дополнительные сведения о ней. В § 17.13, занимаясь проведением многочлена по наименьшим квадратам, мы сначала отбросили тот факт, что кривая должна проходить через начало координат, и в результате нам пришлось пересмотреть всю задачу. Когда эта информация была использована, мы получили более удовлетворительный ответ.

Иногда бывает трудно включить в данные всю известную информацию. Так, в вышеприведенном примере мы знали, что коэффициент при первом члене положителен. Мы не ввели этот факт во входные

данные, но воспользовались им для проверки ответа. Как бы то ни было, прежде чем перейти к следующей стадии работы, полезно ясно представить себе все, что можно заранее узнать про положение дел.

Иногда критический подход к неизвестной ситуации может вызвать другие формулировки задачи, которые в свою очередь приведут к новым идеям. Иногда обнаруживается, что были сделаны излишние ограничительные предположения относительно модели и что их можно легко устранить. Во всяком случае, следует понять роль этих ограничений и включить в вычисления проверки, которые покажут ценность тех или иных предположений. Таким образом, исследование входных данных может вызвать новые требования к содержанию выходной печати.

§ N + 1.4. Обдумывание вычислений

Только после того, как мы ясно поймем, где мы находимся и куда хотим попасть, следует всерьез переходить к вопросу: «Как пройти предполагаемый путь?» Это уже относится к действию книги Поия, все его замечания здесь работают, и мы предполагаем, что читатель знаком с ними.

Хотя задача предложена для решения на вычислительной машине, необходимость машины надо проверить. Аналитическое решение часто гораздо лучше численного, и оценка ошибок иногда может быть сделана более точно, даже когда вычислить эту оценку труднее, чем численно решить саму задачу. Изучение условий задачи может дать одну или больше новых формулировок. Некоторые из них будут просто математическими перефразировками той же задачи, а некоторые, возможно, будут сильно отличаться по существу. Продумывать программу для всех таких возможностей слишком дорого, и придется сразу выбрать некоторые из них. Как правило (из которого опять-таки много исключений), чем ближе математическое утверждение к основным понятиям данной области, тем лучше (предполагается всегда, что все уравнения приведены к безразмерному виду). Замысловатые математические преобразования часто приводят к трудностям при вычислениях.

Принятый план вычислений должен использовать как можно больше первоначальных данных. Математические приближения по формулам должны соответствовать характеру принятой модели. Нужно учесть эффект выборки.

План вычислений должен включать проверки как программирования, так и результатов. Слишком часто на это не обращают должного внимания; мы рекомендуем поэтому ставить себе еще такие вопросы: «как я узнаю, что получил то, что хотел?» и «какие проверки я сделаю или должна сделать машина?» Необходимо, чтобы была вычислена или получена из других источников некоторая из-

быточная информация, чтобы можно было сделать какие-нибудь проверки.

По опыту автора, хороший теоретик может объяснить почти любые полученные результаты, верные или неверные, и по крайней мере может потерять массу времени на выяснение того, верны они или нет.

§ N + 1.5. Повторение предыдущих шагов

Мы все время подразумевали, что указанные этапы работы можно полностью разделить, тогда как на самом деле они часто переплетаются. Тем не менее полезно держать их в уме и повторять сначала, когда на одной из стадий возникают новые данные для другой. Опыт автора показывает, что наиболее распространенная ошибка состоит в поспешности при обдумывании деталей вычисления. Особенно это касается специалиста-вычислителя, так как здесь он чувствует себя как дома и может показать свое искусство. Но все его искусство пропадает зря, если он решает ложную задачу или получает числа, из которых нельзя извлечь ответов на поставленные вопросы.

Трудно, конечно, быть настолько специалистом в частной области, чтобы задавать вопросы о том, нужно ли вычислять то, что просят, или важнее вместо этого вычислить что-нибудь другое. Но существует искусство именно задавать вопросы. Сократ говорил, что он не знает истины, но знает, как задать человеку нужный вопрос, чтобы вытянуть из него истину. Он называл себя повивальной бабкой. Примерно таким же образом специалист-вычислитель должен подходить к заказчику. В конце концов, заказчик должен делать выбор и вести исследование, но разумные предложения со стороны могут помочь прояснить природу его выбора и помочь ему принять решение на многих этапах работы.

§ N + 1.6. Оценка усилий, необходимых для решения задачи

В любой сколько-нибудь развитой науке бывает нужно прикинуть, что произойдет, прежде чем тратить время и деньги. В некотором смысле, чем более развита данная область, тем точнее это можно оценить. Если судить с такой точки зрения, машинная математика находится в крайне элементарном состоянии. Часто доступны лишь самые грубые оценки.

Вот некоторые из обстоятельств, которые надо оценивать:

1. Будут ли влиять ошибки округления, и если да, то насколько?
2. Годится ли взятый интервал?
3. Если имеется итерационный процесс, то сколько примерно потребуется итераций?
4. Сколько времени займут программирование и отладка?

5. Как мы будем проверять результаты, чтобы убедиться, что они верны?

6. Сколько потребуется машинного времени?

7. Когда будут получены окончательные результаты?

Посмотрев на эти вопросы, любой вычислитель скажет, что нынешнее положение дел оставляет желать много лучшего. Нет причин отказываться от попытки делать возможно более реалистичные оценки. Более того, хорошее математическое обеспечение может сильно улучшить оценки по пп. 4 и 7. Наличие диспетчера, систем автоматического программирования, отладочных программ, а также возможность быстро выйти на машину и получать частичные результаты — все это необходимо, если мы хотим иметь точные предсказания по пп. 4 и 7.

В этой книге сделана попытка с разных сторон подойти к ряду таких вопросов. В частности, третья часть книги, по-видимому, дает основание для многих оценок. Однако во многом мы делаем лишь первые пробные шаги в этом направлении, и многое еще предстоит сделать.

Тот, кто стремится стать хорошим специалистом в области вычислительной математики, не должен отмахиваться от этих вопросов как не имеющих ответов, но должен понимать, что умение достаточно точно — не чересчур оптимистично и не чересчур пессимистично — ответить на них — это признак высокого мастерства.

§ N+1.7. Изменения первоначального плана

Почти неизбежно в процессе вычислений появляется новая информация и возникает необходимость вносить изменения в первоначальный план. Но, прежде чем это сделать, надо постараться понять, почему же был выбран неправильный путь. Говорит ли это изменение что-нибудь новое об использованной модели? Надо ли все еще стремиться получить подобные результаты? Не нужно ли по этому поводу вставить новые проверки нашей модели? Нельзя ли что-нибудь понять из самой неудачи или из нового плана?

Изменения не должны вноситься поспешно, им следует посвятить столь же тщательное обсуждение, как и первоначальному плану. Как отмечено выше, если все идет, как задумано, то немного узнаешь. Как раз из неожиданностей могут иногда возникнуть новые вещи. Таким образом, положение, когда приходится изменить первоначальный план, нужно считать скорее счастливой возможностью, чем неудачей. Конечно, если оно возникло из-за беспечности или недомыслия, оно будет лишним примером ценности предварительного обдумывания.

Всегда соблазнительно, втнувшись в задачу, быстро вносить мелкие изменения, не заботясь о последствиях и осложнениях, особенно если результаты обещаны к определенному сроку. И все-таки спешка

в этот момент может свести на нет всю прежнюю тщательную работу. Надо помнить, что «человек — это мастер, а не машина», и хорошей организацией математического обеспечения это может довести до сознания заказчика.

§ N + 1.8. Философия

Из нашего девиза «Цель расчетов — не числа, а понимание», следует, что человек, который должен этого понимания достигнуть, обязан знать, как происходит вычисление. Если он не понимает, что делается, то очень мало вероятно, чтобы он извлек из вычислений что-нибудь ценное. Он видит голые цифры, но их истинное значение может оказаться скрытым в вычислениях.

У Эддингтона есть блестящая история о человеке, который пошел ловить рыбу сетью с ячейками определенного размера. Увидев, что среди пойманных рыб есть самые маленькие, он решил, что это самые маленькие рыбы в море; он допустил ошибку, не учитывая, как происходила ловля рыбы. Так же и при вычислениях; то, что получается, зависит от того, что дано, и от того, что с этим делают. Если не понимать промежуточные процессы, то весьма легко перепутать эффекты использованной при вычислениях модели с эффектами модели, принятой заказчиком при формулировании задачи.

Далее, часто процесс вычисления проливает свет на саму обрабатываемую модель. Вычисление есть средство получения числовых результатов, но это также орудие разума для исследования мира.

На самом деле маловероятно, чтобы большое открытие было сделано профессиональным программистом, стандартным образом программирующим задачи. Если ставится цель понять физическое явление, то автор задачи должен понимать и следить за вычислениями. Это не значит, что он должен выполнять всю мелкую работу, но если он не будет в достаточной степени понимать все, что делает машина, то он вряд ли сумеет извлечь из машины максимум пользы, а также понять смысл даже правильно построенных вычислений.

Опыт показывает, что обычно и легче, и лучше научить специалиста в конкретной области вычислительной математике и программированию, чем наоборот. Но если мы требуем этого от заказчиков, то долг вычислителей приложить все усилия к тому, чтобы уменьшить для них трудности обучения. Произвольные правила, особый жаргон, бессмысленный формализм, изменения в методах и обозначениях, препятствия в получении времени — все должно быть сведено до минимума. Особенно внимательно надо следить, чтобы этот груз трудностей не по существу не возрастал с появлением новой машины.

Наука о том, как вместо просто быстрое действия машины использовать численные методы и библиотечные программы, переживает

период детства и является одной из наиболее важных областей исследования в будущем. Работа в этой области требует опыта по использованию машин в каждодневной работе. Развивать ее, безусловно, стоит.

§ IV + 1.9. Заключительные замечания

Для прогресса машинной математики очень важно, чтобы интуитивные методы, которыми мы теперь пользуемся, были более ясно поняты и приведены, насколько возможно, к явным и удобным для вычислителей рекомендациям.

Нет необходимости напоминать читателю, что большинство предыдущих рекомендаций и замечаний представляют личное мнение, выработанное автором в отдельной лаборатории, и что они вовсе не обязательно всегда применимы. В их защиту можно сказать, что они опираются столько же на здравый смысл, сколько на опыт. Если читателю они не понравятся, то пусть он не спорит на эту тему, а изложит свои собственные соображения.

1. Adams E. P. and Hippisley R. L., Mathematical Formulae and Tables of Elliptic Functions, Publication 2672, Smithsonian Institution, Washington, D. C., 1947.
2. Blackman R. B. and Tuckey J. W., The Measurement of Power Spectra, Dover Publications, New York, 1959.
3. Boole George, Treatise on the Calculus of Finite Differences, republished by Stechert-Hafner, Inc., New York, 1946.
4. Bromwich T. J. J'a, An Introduction to the Theory of Infinite Series, 2d ed., Macmillan & Co., Ltd., London, 1947.
5. Buckingham R. A., Numerical Methods, Sir Isaac Pitman & Sons, Ltd., London, 1957.
6. Campbell G. A. and Foster R. M., Fourier Integrals for Practical Applications, D. Van Nostrand Company, Inc., Princeton, N. J., 1948.
7. Carslaw H. S., Fourier Series and Integrals, Dover Publications, New York, 1930.
8. Erdélyi A. (ed.), Tables of Integral Transforms, vol. 1, McGraw — Hill Book Company, Inc., New York, 1954.
9. Вазов В. и Форсайт Дж., Разностные методы решения дифференциальных уравнений в частных производных, ИЛ, М., 1963.
10. Fox L., Mathematical Tables, vol. 1, Her Majesty's Stationery Office, London, 1956.
11. Fox L., Numerical Solution of Two-point Boundary Problems, Oxford University Press, London, 1957.
12. Hartree D. R., Numerical Analysis, 2d ed., Oxford University Press, London, 1958.
13. Hastings Cecil, Jr., Approximations for Digital Computers, Princeton University Press, Princeton, N. J., 1955.
14. Hildebrand F. B., Introduction to Numerical Analysis, McGraw — Hill Book Company, Inc., New York, 1956.
15. Hodgman C. D. (ed.), Handbook of Physics and Chemistry, Chemical Rubber Publishing Co., Cleveland, 1960.
16. Хаусхолдер А. С., Основы численного анализа, ИЛ, М., 1956.
17. Джексон Д., Ряды Фурье и ортогональные полиномы, ИЛ, М., 1948.
18. Jolley L. B. W., Summation of Series, Chapman & Hall, Ltd., London, 1925, and Dover Publications, New York, 1960.
19. Jordan Charles, Calculus of Finite Differences, Chelsea Publishing Company, New York, 1947.
20. Kopal L., Numerical Analysis, John Wiley & Sons, Inc., New York, 1955.
21. Kuntzmann J., Méthodes numériques, Interpolation — dérivées, Dunod, Paris, 1959.
22. Кунц К., Численный анализ. Техника, Киев, 1964.
23. Ланцош К., Практические методы прикладного анализа, Физматгиз, М., 1961.

24. Lighthill M. J., *Fourier Analysis and Generalized Functions*, Cambridge University Press, London, 1958.
25. Милн В. Э., *Численный анализ*, ИЛ, М., 1951.
26. Милн В. Э., *Численное решение дифференциальных уравнений*, ИЛ, М., 1954.
27. Milne-Thomson L. M., *The Calculus of Finite Differences*, Macmillan & Co., Ltd., London, 1933.
28. Mineur H., *Techniques de calcul numérique*, Librairie Polytechnique, Paris, 1952.
29. *Modern Computing Methods*, 2d ed., Philosophical Library, Inc., New York, 1961.
30. Muir T., *A Treatise on the Theory of Determinants*, Dover Publications, New York, 1960.
31. National Bureau of Standards, *Tables of Functions and Zeros of Functions*, Applied Mathematics Series, № 37, Washington, 1954.
32. National Bureau of Standards, *Tables of Lagrangian Coefficients*, Washington, Columbia University Press, New York, 1944.
33. Nörlund N. E., *Vorlesungen über Differenzenrechnung*, Springer — Verlag, Berlin, Vienna, 1924.
34. Островский А. М., *Решение уравнений и систем уравнений*, ИЛ, М., 1963.
35. Pairman Eleanor, *Tracts for Computers No. 1, Tables of the Digamma and Trigamma Functions*, Cambridge University Press, London, 1919.
36. Пойа Д., *Как решать задачу*, Учгедгиз, М., 1959.
37. Ralston A. and Wilf H. S., *Mathematical Methods for Digital Computers*, John Wiley & Sons, Inc., New York, 1960.
38. Скарборо Дж., *Численные методы математического анализа*, ГТТИ, М., 1934.
39. Стефенсен И. Ф., *Теория интерполяции*, ГТТИ, М., 1935.
40. Трантер К. Дж., *Интегральные преобразования в математической физике*, ГИТТЛ, М., 1956.
41. Ватсон Г. Н., *Теория бесселевых функций*, ч. I, ИЛ, М., 1949.
42. Уиттекер Э. Т. и Робинсон В., *Математическая обработка результатов наблюдений*, ГТТИ, М., 1935.
43. Уиттекер Э. Т. и Ватсон Г. Н., *Курс современного анализа*, тт. I, II, Физматгиз, М., 1963.
44. Зигмунд А., *Тригонометрические ряды*, тт. I, II, «Мир», М., 1965.